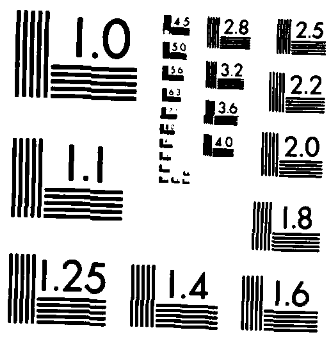
AD-A154 862 INVESTIGATION OF DBMS (DATA BASE MANAGEMENT SYSTEMS) FOR USE IN A RESEARCH ENVIRONMENT(U) RAND CORP SANTA MONICA CA P N ROSENFELD FEB 85 RAND/P-7002 1/2

UNCLASSIFIED F/G 9/2 NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(2)

INVESTIGATION OF DBMS FOR USE IN A RESEARCH ENVIRONMENT

Pilar N. Rosenfeld

February 1985

P-7002

The Rand Paper Series

The Rand Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90406-2138

PREFACE

This paper is a thesis submitted to satisfy requirements for the degree of Master of Science in Computer Science at California State University, Northridge. The oral defense was presented and accepted on July 6, 1984. The text processing and computer costs were supported by the Rand Educational Program and the Rand Computer Services Directed Study program.

The Rand Corporation was used as a case study for an investigation of Data Base Management Systems (DBMS) for use in a research environment. This paper should be of interest to DBMS designers and researchers interested in another data management alternative.

ACKNOWLEDGMENTS

This masters thesis is dedicated to my husband Gary Rosenfeld, and my parents Eleodoro and Felicitas Montes who have always encouraged me to pursue my goals.

I wish to express my appreciation for the continuing support and suggestions I received from the members of my masters committee: Peter Smith (chairman), Suzanne Polich, and Michael Barnes. I am also grateful to the many people at The Rand Corporation who helped me throughout the project, which includes participants in the DBMS feature survey.

ABSTRACT

AN INVESTIGATION OF DBMS FOR USE IN

A RESEARCH ENVIRONMENT

by

Pilar N. Rosenfeld

Master of Science in Computer Science

This report is an investigative study on whether a DBMS has a place in a research environment. The study concentrated on the use of large social science data sets. The following topics were examined:

- how social science data sets are used in a research environment
- the data usage and needs of an existing research institution (The Rand Corporation)
- the differences between research and business applications
- the possible DBMS configurations within a research environment
- the opinions of Rand computer users when rating the importance of DBMS features
- evaluation of commercial DBMS for use in a research environment.

Conventional data base management systems ~~(DBMS)~~ have been very successful with business/corporate data bases, but DBMS are not widely used with research data bases. There are significant differences between the business and research data management needs. These include different retrieval and update specifications, the need for statistical routines, and less financial data base support. Much of research analysis requires the use of statistical procedures. Hence, a DBMS configuration within a research environment must include some access to statistical procedures. Given these requirements, there are a few commercial DBMS which could be considered for a research environment.

CONTENTS

PREFACE iii

ACKNOWLDGMENTS v

ABSTRACT vii

FIGURES xi

TABLES xiii

Section

I. Introduction 1

II. Data Base Management System Definitions and Terminology ... 3

III. The Research Environment 8

THE RESEARCH PROCESS 9
Detailed description of the research process 10

IV. A Research Institution (The Rand Corporation) 17

Description of the Rand Corporation 17
RAND COMPUTING PROFILE 19
DATA MANAGEMENT METHODS USED AT RAND 24

V. BUSINESS VS RESEARCH 31

VI. Research Environment DBMS Requirements 40

Data Base Management System Configurations 40
Important DBMS Features Survey Description 45

VII. Evaluation of Commercial Data Base Management Systems 56

VIII. Introduction of DBMS into a research environment 65

IX. Summary and Conclusions 69

X. Glossary 76

Appendix A

Example of Social Science Data Bases 78

Appendix B

Description of FY82 Project Leaders' Computing Profile 80

Appendix C

Example of PL/I Program Used for Data Management 81

Appendix D

Example of a RETRO Program 86

Appendix E The Statistical Analysis System (SAS) 87

Appendix F Survey of the Importance of Data Management Features . 89

Appendix G

Cover Letters used with DBMS Survey 92

Bibliography 94

FIGURES

1. Master File Stage of Research Process 13
2. Analysis File Stage of Research Process 15
3. DBMS Interface with Statistical Software Package 44

TABLES

1. FY82 Computer Expenditures by Divisions 20

2. Computing Activities at Rand 21

3. Software Application Packages of Future Importance 23

4. Data Management and Statistical Analysis of Numeric Data ... 25

5. Rating the Importance of DBMS Features 51

6. Top Ranked DBMS Features

   The Computer Services Department Only 53

7. Top Ranked DBMS Features

   The Research Staff Only 54

8. Top Ranked DBMS Features

   The Computer Services and Research Staff 55

9. Commercial DBMS Evaluated for Use in a Research Environment
   56

10. Comparison of Data Base Management Systems 58

11. Commercial DBMS: Five Year License Fee Comparison 64

I. INTRODUCTION

Currently there is no standard method of managing large social science data files. Each research project has its own method of accessing data. Moving to a new project requires learning not only the new data and analysis, but also the data organization and retrieval conventions. The conventions used to organize the data are determined by the initial data manager or programmer; the data management technique is limited to that person's experience and preferences. Work can be difficult to transfer to another programmer, and very few programs can benefit more than one project.

The use of DBMS appears to be a natural solution to the problem of managing large data files. However, many researchers feel that current commercial data base management systems (DBMS) are not suited to their needs, even though DBMS have been successful and accepted by the corporate/business community. Thus, an investigation has been conducted into the use of data base management systems for large social science data bases in a research environment. The purpose in addressing this issue is to describe the "research environment" and identify the particular data base management requirements. The information provided might prove useful to DBMS designers, and researchers interested in better data management.

This document reviews the major components of the study. The next section (section II) presents an overview of DBMS terminology and concepts. This is followed by a thorough description of the research environment and the uses of social science data bases (sections III and IV). Section III provides a general description of a typical research environment, while section IV uses an existing research institution, The Rand Corporation, as an example. A comparison of the research environment and the business environment is presented in section V. This section will identify why DBMS are successful in business and not in research. Section VI covers the interrelationship between the research environment and the DBMS in two ways. First, the possible DBMS configurations within a research environment are described. Second, the

opinions of the Rand computer users on the important DBMS features are presented. Given the research data base management requirements established in previous sections, section VII presents an evaluation of commercial DBMS. Section VIII is a brief discussion on introducing a DBMS into a research environment. The final section (section IX) is a summary of the entire study.

II. DATA BASE MANAGEMENT SYSTEM DEFINITIONS AND TERMINOLOGY

Data base terminology and data base theory will be briefly presented in this section. To facilitate this discussion, it is necessary to set some common definitions. They are: *Data* is a group of non-random symbols that represent quantities, actions, things, facts, concepts or instructions in a way suitable for communication and processing by humans or machines. *Information* is data that has been processed and presented. A *record* is a group of related data items. A *file* (data set) is a collection of related records. A *data base* is a collection of files logically related in such a way as to improve access to the data and minimize redundancy of data. A *data base management system* is a set of programs that function to create and update the data base, retrieve data and generate reports from the data base. A *conceptual model* (data model) is a representation of the information content of the data base.[1]

To get a better understanding of the data base and data base management we now consider four questions:

1. *What is a data base?*

James Martin, a well known author in the data base field, defined a data base as

> a collection of interrelated data stored together without unne redundancy to serve multiple applications. The data are ed so that they are independent of the programs which use m. A common and controlled approach is used in adding new ta and modifying and retrieving existing data. The data structured so as to provide a foundation for future a plication development.[2]

[1] Mohammed H. Omar, "DBMS Simplified," *Data Management*, October 1982, p. 26.

[2] James Martin, *Computer Data Base Organization* (Englewood Cliffs, New Jersey: Prentice-Hall, 1977), p. 22.

IV. A RESEARCH INSTITUTION (THE RAND CORPORATION)

The previous section described the research environment. A specific research institution was studied, to establish realistic data management needs and expectations. The Rand Corporation based in Santa Monica, California was used as the case study. The emphasis was on the computer application of large social science data bases. First, a general description is given of Rand, the nature of the work done at Rand, and the computer facilities available. To get more insight into the Rand computing profile, the results from a survey given to project leaders on computing activities is summarized. The final portion of this section deals with the current data management methods used at Rand. This will give the reader some perspective on what methods are currently used and how a DBMS might assist in the data management effort.

DESCRIPTION OF THE RAND CORPORATION

The Rand Corporation is a private nonprofit institution engaged in research and analysis of matters affecting national security and the public welfare. Rand conducts its work with support from federal, state, and local governments; from private foundations and other philanthropic sources; and from its own funds drawn from fees earned. The work done at Rand is divided in two divisions: National Security Research and Domestic Research. There are some social science data bases under the National Security division, on recruiting and enlistment; however most social science data bases come under the Domestic Research division. Within the Domestic Research division there are several research programs which specialize in particular topics. The research programs include: Criminal Justice, Educational and Human Resources, Energy Policy, Health Sciences, Housing and Urban Policy, Labor and Population, and Regulatory Policies and Institutions.[1]

---

[1] "Rand Annual Report 1982-1983," (Santa Monica, California: The Rand Corporation, 1983).

be made. There is also a tremendous expense in conducting a survey (experiment) and collecting the original data.

The difficulty of building the data analysis file (the data transformation step) is dependent on the complexity of the data relationships inherent to the data. It can also be influenced by how the data is arranged in the master file (e.g., whether the data is already integrated). There are times when it is advantageous to rebuild the master file into a more suitable master working file.

DBMS Function within the Research Process

We have an understanding of the research process; we will now see how a DBMS might contribute to this process. The majority of the DBMS functions involve the master file stage. The DBMS will organize the master file. The data description, data editing, data transformation and data modification steps are all done through the DBMS. The purpose of the analysis file stage is to prepare data for analysis and perform statistical analysis on that file. The DBMS aids in the data transformation and data modification steps of the analysis file stage, by facilitating this process. Since DBMS rarely have statistical procedures available, the analysis file must use other software. A DBMS with an interface to statistical software would be highly desirable. This is a fairly simplistic view of the DBMS place in the research environment, and will be elaborated further in the configuration portion of Section VI.

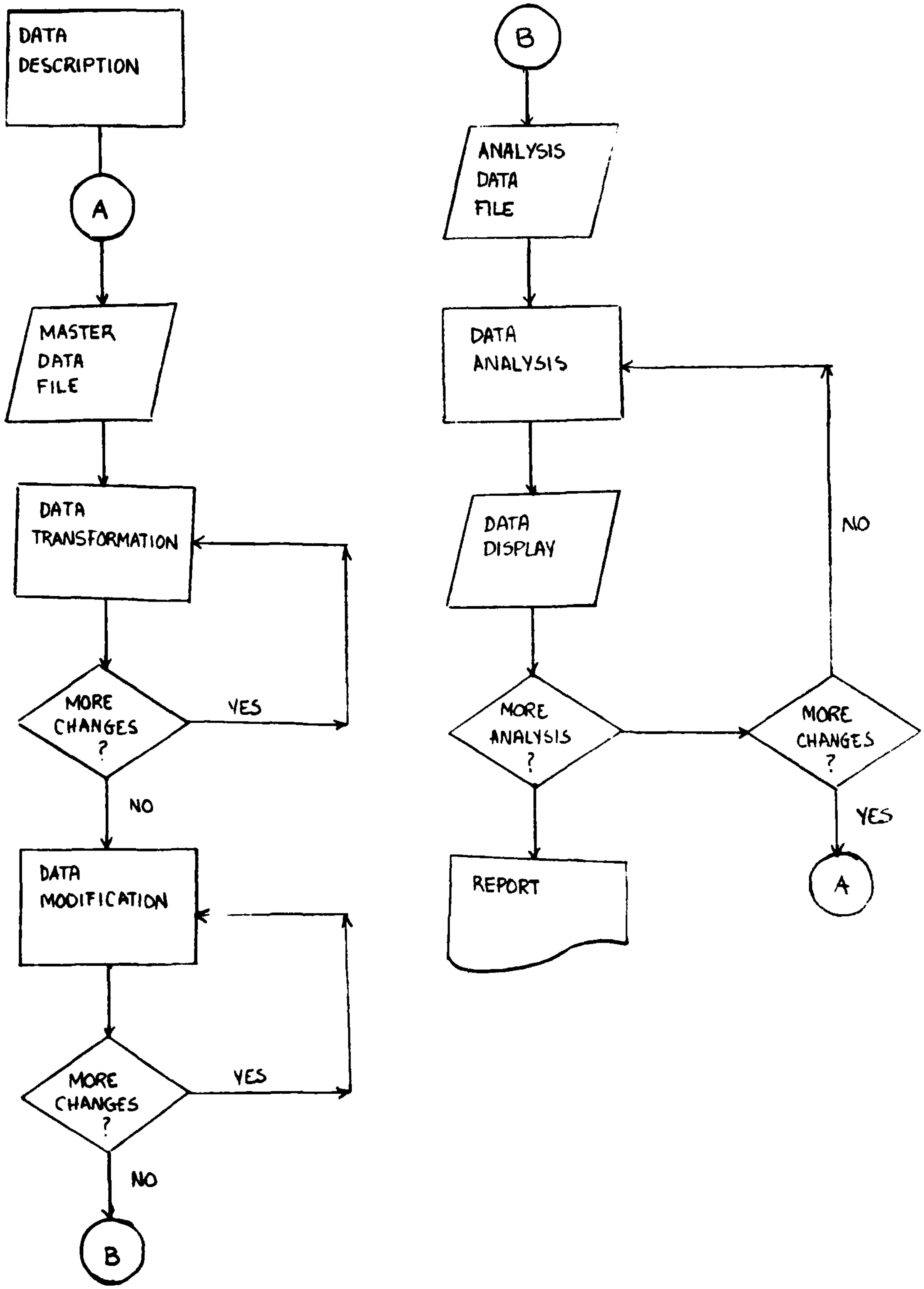


Fig. 2 -- Analysis File Stage of Research Process

disaggregation. The more master files used, the more complex the transformation procedure becomes, especially if the files need to be restructured. As mentioned in the master file stage, the unit of analysis (record key) must remain consistent.

3. *Data modification*: many derived variables are built in the master file stage. The researcher tries to keep only the relevant (most important to the study) information.
4. *Data analysis*: some of the objectives of data analysis are the analysis of measurement error (e.g., reliability and validity analysis), exploratory (e.g., factor analysis and stem and leaf analysis), descriptive (e.g., tabulation), explanatory (e.g. linear models) [3]. The analysis step involves running statistical procedures on the analysis file. There are statistical software packages which provide most of the statistical procedures needed.
5. *Data display*: data display consists of summary reports, tables and graphic displays of the data and analysis results.[4]

The steps in the analysis file stage are illustrated in figure 2. The analysis phase can be investigative in nature, therefore it is usually difficult to determine all the analysis file specifications. This is why it is sometimes necessary to go from the data analysis step back to the master files to pick up values. Note that some of the steps are the same as those in the master file stage; they have the same type of processes but have different goals.

A researcher may not be concerned with the entire research procedure. In fact the majority will start with the analysis file processing. This is called secondary analysis. Secondary analysis is done because data bases are rich in information. Thus, many studies can

---

[3] Ronald E. Anderson and Francis M. Sims, "Data Management and Statistical Analysis in Social Science Computing," *American Behavioral Scientist*, Vol. 20 No. 3, January/February 1977, p. 394.

[4] Ronald E. Anderson and Francis M. Sims, "Data Management and Statistical Analysis in Social Science Computing," *American Behavioral Scientist*, Vol. 20 No. 3, January/February 1977, p. 371.

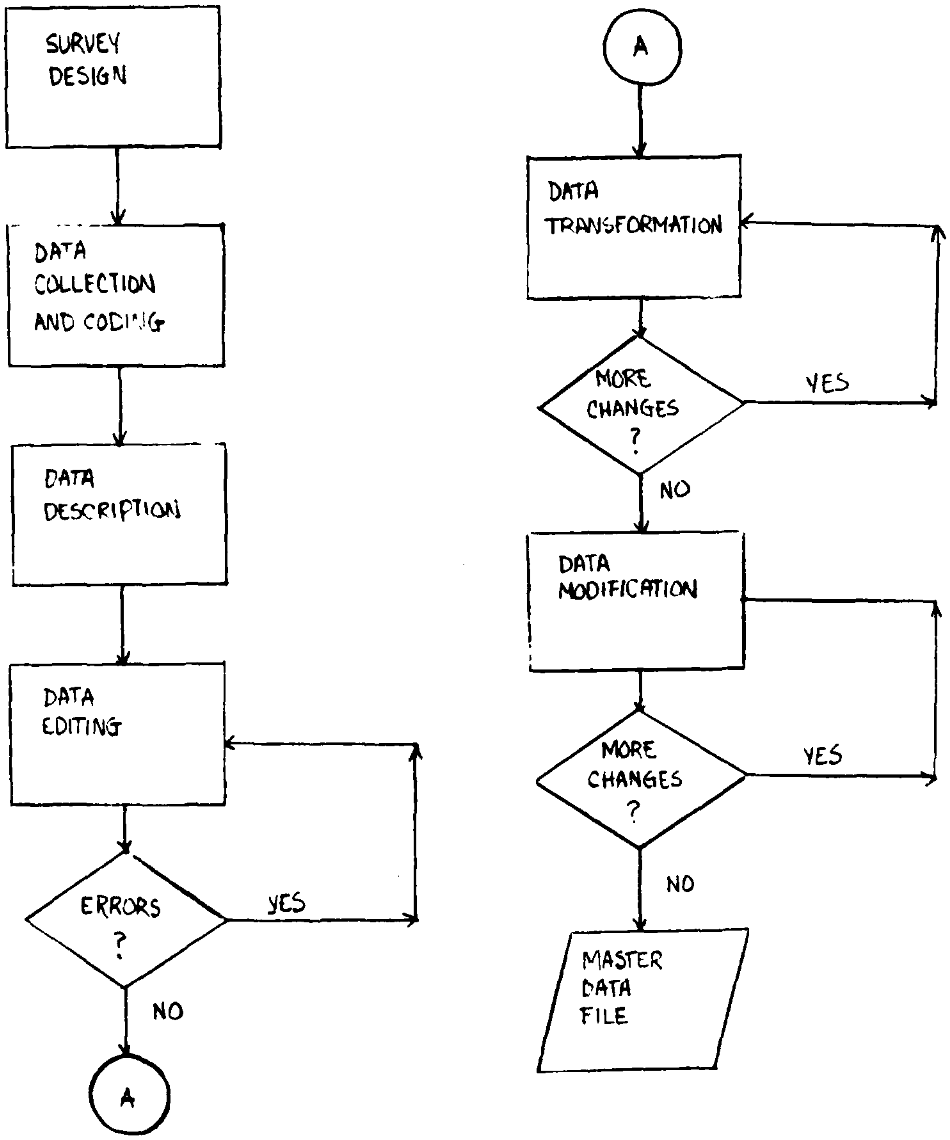


Fig. 1 -- Master File Stage of Research Process

Figure 1 shows the steps used in the master file stage of the research process, from the initial design of the survey to building the master file. There are two situations in which a master file is created: (1) a contract has been made to conduct the survey and build the master file. The entire process is done in this case. (2) The master file has been received from another source and was not carefully prepared or is in a form that is difficult to use. In this case the master file is rebuilt, starting with the data description step.

The Analysis File Stage

The analysis file stage of the research process has the following five steps:

1. *Data description*: the researcher has a hypothesis to test. First he/she must decide what "unit of analysis" to use, for example, a study by individual or family, or by migrant or migration. The "unit of analysis" is analogous to the record key; there is one record per unit of analysis. For example the unit of analysis for the Living Arrangements project is the young adult and the analysis file has one record per young adult. Next the researcher determines which variables will enable the test of the hypothesis. Data sources (master files) are examined to see which supply the desired variables. The data description is the blueprint for the analysis file, giving the record structure (the unit of analysis), the master files to use and the variables to be included. Another consideration in developing the data description is that the resulting analysis file must be compatible with the computer programs that will perform the statistical analysis. Most statistical analysis procedures require flat fixed format files.
2. *Data transformation*: the analysis file should be organized with one record per unit of analysis. The data transformations used to achieve this are subsetting, merging, aggregation, and

---

Statistical Analysis in Social Science Computing," *American Behavioral Scientist*, Vol. 20 No. 3, January/February 1977, p. 371.

2. *Data collection and coding*: conducting the experiment and collecting the data. Coding or transcribing the data in a computer compatible form.
3. *Data description*: defining the file structure and the data element format of the survey data to build the master file. In the master file stage, the data is usually organized like the survey questionnaire, the record unit is the same as the unit in the survey, such as family unit, and the data is ordered according to the position on the questionnaire. Many times a codebook is prepared, which includes a description and location of the variables, the possible values (with translations for coded values), and initial frequencies. The data description step prescribes the subsequent data transformation step.
4. *Data editing*: this is the process of "cleaning" the data, which means detecting and correcting errors. No matter how thoroughly the collection and data coding/transcribing was done, there are always errors in the data. Data editing is a critical step to insure the accuracy of the data and the resulting analysis.
5. *Data transformation*: data transformation is defined on the record structure level and is involved in organizing the records. Most master files are organized by the unit in the survey questionnaire; therefore there is usually little data transformation required. What is important is keeping the record key (unit) consistent. Some master files are built from several questionnaires, then the transformation may include merging, aggregation (the unit is lower and must be summ- ized), and disaggregation (the unit is higher and must be apportioned to each record unit).
6. *Data modification*: data modification is defined on the data element (variable) level. The data element is recoded or derived from other variables, creating a "derived variable." An example might be that given the education level for each parent, the average education level for both parents is built.[2]

[2] Ronald E. Anderson and Francis M. Sims, "Data Management and

Computer Processing in the Research Process

Computer processing is used for the statistical analysis and related data management. The analysis is done by running statistical procedures on the data. Data management prepares the data for this analysis. This involves data base planning, data entry, data manipulation, and data display. The "research process" steps corresponding to the computer processing are steps 4-6 (noted above). Within these steps the social science data base goes through two stages.

1. The Master File. The master (original data) file is a collection of surveyed data, put into a computer readable form.
2. The Analysis File. The analysis (working) file is built from master files and is the actual file used by the researchers in their analysis.

In the example of the "Living Arrangements and Family Formation of Young Adults" project, step 4 is skipped because "master files" were available. The project uses three longitudinal data sets, the major data set is the Parnes National Longitudinal Surveys of Young Women and Young Men, described in Appendix A. Step 5 consists of building the analysis file from the master files. In step 6, a researcher performs statistical analysis on the analysis file.

DETAILED DESCRIPTION OF THE RESEARCH PROCESS

A detailed description of the research process will be given in terms of the master file stage and the analysis file stage.

The Master File Stage

The master file stage is concerned with building the master file, and includes the following six steps:

1. *Survey design*: the actual design of the survey or experiment. The survey questionnaire and instructions will affect how easy the resulting data base will be to use.

An example of a research study is the "Living Arrangements and Family Formation of Young Adults" project. The hypothesis is that experience in nonfamily living during young adulthood affects the timing and nature of subsequent family formation. The three central questions addressed are (1) Young adults who have lived in nonfamily households more likely than others to postpone or even forego marriage? (2) Does the experience of nonfamily living before marriage affect the likelihood and timing of a first birth, attitudes toward family size, number of children expected, and number of children born? (3) Does the experience of premarital nonfamily living affect the stability of subsequent marriages?[1]

THE RESEARCH PROCESS

Once a social science researcher has a hypothesis, such as described above, how does he/she arrive at the results? The procedure which gets the researcher from the hypothesis to the final results, is called the "research process." A simplified view of the social science research process is:

1. State the hypothesis.
2. Determine what data will be needed to test the hypothesis.
3. Examine existing data sources for the desired data.
4. If there are no data sources with the appropriate data, then a data base must be collected.
    a. design the survey
    b. collect the data
    c. put the survey in computer readable form (organize and edit)
5. Prepare the data for analysis
6. Perform the analysis
7. Report the results

[1] Linda J. Waite, *Changing Living Arrangements and Family Formation of Young Adults*, Proposal submitted to National Institute of Child Health and Human Development, NICHD-SBS-81-3, February 1981.

III. THE RESEARCH ENVIRONMENT

In this section we give an overview of the application of social science data bases. The social science environment is briefly described in terms of what a social science data base is, where social science research is done, who does the research, what is produced from the research, and how the social science data is used within the "research process." First the research process is briefly defined. Then the steps concerned with the computer processing of the social science data bases are described in more detail.

Social science data can be defined by its content and by the way in which the data is used. Examples of social science data are population or income counts and health and medical statistics. Usually the data is derived from surveys and censuses. Some surveys are retrospective, that is, questions are asked about events in the past (e.g., past illness). Another type of survey is longitudinal, in which a set of questions are asked repeatedly over a period of time (e.g., given each year for a seven year period). Examples of typical social science data bases are given in Appendix A. Social science data are analyzed using statistical methods. Some of these methods include marginals, cross tabulations, and linear regressions.

The Research Environment

Social science data bases are analyzed in a "research environment." Social science research environments exist in research institutions and educational institutions. The analysts or researchers have varied academic training, including economics, sociology, demography, psychology, anthropology, and nutritional epidemiology. These researchers use econometric and statistical procedures in the analysis of data. The results of the analysis are presented to professional audiences through academic presentations, papers given at scientific meetings, and publications in scholarly journals. The results can also be presented to policy oriented audiences through Congressional testimony, policy briefings, and publications in policy journals.

To allow data independence such that the programs accessing the data base are relatively independent of the storage and access methods.[6]

---

[6] J. P. Fry, and E. H. Sibley, "Evolution of Data Base Management Systems," *Computing Surveys*, Vol. 8, March 1976, p. 8.

Data base query languages allow the user to add, retrieve, update, and display data from a data base. Usually query systems are written so that noncomputer specialists can use them. There are also other types of data base utility software. These include the following: backup/recovery, password security, DB/DC (Data base/data communications), inscription/description software, image management software (text, graphics), audit trail utilities, data base tuning utilities, data base development aids, data base reloading and reorganizing aids, and data base sizing and responsiveness aids.[5]

3. *What are the different data base management system designs?*

Data base management systems usually differ in the type of view (conceptual model) allowed. There are three categories of conceptual models: hierarchical, network, and relational. In the hierarchical model, the structure is in the form of a tree. The hierarchical model is a subset of the network model. In the hierarchical data base, an individual record (node) may have only one owner (parent). In the network data base, one record type may be owned by any number of record types. The network is represented as a plex (graph). Relationships between fields are established via links. The relational model is based on "tables." A table is called a relation. A column is called an attribute. The major appeal of the relational approach is its simplicity: tables are a simple and natural way of viewing data.

4. *What are the objectives of data base technology?*

The primary objectives of data base organization were summarized by Fry and Sibley:

> To make an integrated collection of data available to a wide variety of users
>
> To provide for quality and integrity of data
>
> To provide privacy and security measures

[5] Daniel S. Appleton, "Implementing Data Management," *National Computer Conference Proceedings*, Vol. 49, 1980, p. 314.

Usually the more elaborate DBMS have all of these tools available, but they may be individually priced or priced in groups. Less elaborate DBMS (such as those traditionally found on microcomputers) have only some of these facilities.

A data dictionary is a list of all the elements contained in the data base and the relationships which are established among these elements. A data dictionary describes each element, its synonyms, the organization responsible for updating, the data format, its security requirements and a description of what the data element means. The data directory supplements the data dictionary. Its function is to describe how each individual data element is used and where it is used.

Data definition languages (DDL) describe the content and structure of both the "schema" and the "subschema." A schema is the description of a complete logical data base, and a subschema is the description of a subset of that data base which is utilized by an individual computer program. The DDL are oriented toward describing the data in terms of the conceptual model provided by the DBMS.

Data manipulation languages (DML) allow the data base to be accessed by host language (FORTRAN, COBOL, PL/I) programs. Typical capabilities include:

Create: Create an occurrence of a data record.

Store: Store data in the data base.

Fetch: Retrieve data from the data base.

Insert: Put a data record in a data relationship.

Modify: Change data in a data record.

Find: Locate a data occurrence.

Delete: Delete a data record.

2. *What constitutes a data base management system?*

A data base management system (DBMS) simply put, is the software that supports "data bases." There is no standard definition of what constitutes a DBMS, other than being an integrated, complementary set of tools/features to handle data base requirements. The fundamental features provided by a DBMS are:

Centralized control of the data, which implies reduction of data redundancy, shared data, protection of the data base integrity, and enforcement of standards.

Data independence; the ability to modify the data base structure without having to modify the programs which use the data.

Provision for complex file structures and access paths, such that relevant relationships between data units can be expressed in a more natural form.

Generalized facilities for storage, modification, reorganization, analysis and retrieval of data. This is done by interface with programming languages, a user query language, or both.

Security to prevent unauthorized access to stored data.

Mechanisms to enable easy recovery or restoration of the data base.[3]

The DBMS composition is dependent on the vendor. Some vendors provide more tools/features than others. Basic data management tools include:

Data Dictionaries/Directories

Data Base Definition Languages

Data Base Manipulation Languages

Data Base Query Languages

Data Base Utilities (Control)[4]

---

[3] Ian Palmer, *Data Base Systems: A Practical Reference* (Wellesley, Massachusetts: Q.E.D. Information Sciences Inc., 1975) p. 1-3.

[4] Daniel S. Appleton, "Implementing Data Management," *National Computer Conference Proceedings*, Vol. 49, 1980, p. 314.

The Nature of Domestic Research at Rand

Rand literature describes Rand domestic research as follows:

> The hallmark of Rand research is that it examines systematically as many relevant aspects of a problem in as large a context as possible. Thus a typical Rand study is addressed to policymakers. It identifies and evaluates alternative solutions or courses of action, taking into account their costs and benefits and their sensitivity to future developments - features that may or may not be part of current assumptions.[2]

Most of the research projects can be categorized as "systems analysis." In particular the social science systems includes the educational system, the criminal and civil justice systems, the health-care system, federal agencies' management practices, the welfare system, and systems for introducing social change. An example is the study of school governance and finance in the context of judicial reform and changing economic, political, and demographic factors. The research projects can also be described by the perspective or technique used. The techniques used on social science data bases include, evaluations, social experiment, study design, modeling, and literature review and synthesis.

The *evaluation* technique is used to test procedures and examine results of a project to determine whether its objectives could be achieved more effectively in other ways. A *social experiment*, a relatively new technique, is analogous to lab experiments except this experiment is done in a social environment where people are the subjects and social issues are the hypotheses being tested. *Study design* is a preliminary examination of the rationale and analytical methods to be used on a proposed research project. *Modeling* is a way of mathematically describing a system. *Literature review and synthesis* is the study and examination of extensive but fragmented literature for the purpose of deriving specific information.

[2] "Domestic Research at Rand," (Santa Monica, California: The Rand Corporation, January 1981).

Computer Facilities at Rand

The major computer at Rand is an IBM 370/3032 mainframe with the OS/MVS operating system. Programs are edited and executed with WYLBUR (an interactive batch system) and TSO. The computer programming languages maintained are FORTRAN, PL/I, BASIC, APL, PASCAL, COBOL, and the assembly language BAL. Rand maintains several statistical software packages: SAS (Statistical Analysis System), STATLIB (A Statistical Computing Library), SPSS (Statistical Package for the Social Sciences), BMDP, LIMDEP, and LISREL. There is a data management system for performing print and extraction operations on large data sets (PRTEXT). The SIR (Scientific Information Retrieval) data base management system is also available, because Rand was the test facility for the IBM version of SIR. Since there are a variety of computers at Rand, there is software that links the IBM mainframe, a PDP, several VAXes and several IBM PCs for the purpose of data transfers.

RAND COMPUTING PROFILE

To provide an understanding of the Rand computing profile, the computer utilization was examined. The source of this information is the "FY82 Project Leaders' Computing Profile Survey" (FY82 is fiscal year 1982).[3] A description of the survey is given in Appendix B. Only the topics that give a general overview of the Rand computing profile, or which relate to data management and statistical analysis are presented here. These include the following: computer expenditures, computer activities, data storage, size of files, software application packages and languages, data management and statistical analysis.[4]

---

[3] Donald P. Trees, *Results from the FY82 Project Leaders' Computing Profile Survey:* Service Satisfaction, Future Resource Requirements and Policy Issues, (Santa Monica, California: The Rand Corporation, June 1983).

[4] Note the information for data storage and size of files has two sample sizes. Past/current computer usage only included personally interviewed respondents (the number of respondents, N = 41). Information concerning future expectations included both the interview and mail respondents (N = 115).

Computer Expenditures

The computer expenditures information was divided into three Rand divisions. As discussed earlier in this section, there are the Domestic and National Security divisions. The third division included in the Computing Profile Survey was the Corporate division, which is involved in the Rand "business" applications. The Domestic division accounted for the majority of the computer expenditures for FY82. Table 1 shows the breakdown of computer expenditures by division.

Computer Activities

The Project Leader Survey asked the respondents which computer activities were done by their project. At least seventy percent of Rand projects perform data and file maintenance, text processing, and numerical analysis or mathematical modeling. Other major activities (greater than forty percent) are interactive data retrieval and updating, simulation, repetitive numeric data entry from terminals, and graphics. Table 2, "Computing Activities at Rand" presents an overview of current and future estimates for computer activities.[5]

Table 1

FY82 COMPUTER EXPENDITURES BY DIVISIONS

division	all projects
domestic	57%
national security	26%
corporate	17%

The major computer used was the IBM 370/3032. This computer expenditures table excludes text processing expenditures. Text processing accounted for approximately sixteen percent of the total

[5] Donald P. Trees, *Results from the FY82 Project Leaders' Computing Profile Survey:* Service Satisfaction, Future Resource Requirements and Policy Issues, (Santa Monica, California: The Rand Corporation, June 1983).

computing revenues, and was done on the PDP and VAX machines.

Table 2

COMPUTING ACTIVITIES AT RAND

(Percentage of projects which did the following activities)

Computing Activity	Done FY82 Projects	Expected Future Increase
Text Processing	*80%	50%
Statistical Analysis	*79%	Not asked
Graphics	42%	54%
Interactive Statistical Analysis	23%	37%
Numerical analysis or mathematical modeling	70%	36%
Simulation	45%	30%
Data and file maintenance	87%	34%
Interactive data retrieval and updating	46%	27%
Questionnaire or forms development	25%	15%
Repetitive numeric data entry from terminals	43%	16%
Textual or bibliographic data analysis	18%	20%
On-line budget preparation and cost monitoring	5%	14%
Geographic coding and matching	17%	11%
Electronic Mail	*27%	*30%

* Estimated from interview data only (N=41)
All respondents, interview and mail data (N=115)

Data Storage

The data storage information includes both text and numerical data files. There was about a fifty/fifty split between storage on disk and tape. Twenty-one percent of the users had multi-volume tape data files, where the largest data file was three volumes. The average for the greatest number of tape drives used at one time was two. Thirty-seven percent of the projects utilized data which was considered sensitive and subject to Department of Defense, proprietary or privacy regulations and restrictions.

Size of Files

The size of file information only includes numeric or statistical data, as opposed to text data. Of the large and medium users, ninety percent of the projects utilized numeric data files. The survey collected data on the size of the largest numeric data file. The average size was 63,624 observations with a logical record length of 1,100 bytes. The average number of original (master) files per project was fourteen. Sixty-six percent of the time, the master files were larger than the analysis files. Thirty-one percent of all respondents expect the size of the data files to increase, and thirty-six percent expect the number of master files to increase. Fifty-five percent expect to acquire data files from Federal agencies and data archives. Of these, forty-five percent expect an increase in the number of files acquired.

Software Application Packages and Languages

The software application packages used the most at Rand are SAS (Statistical Analysis System) and STATLIB (A Statistical Computing Library). The survey asked project leaders which software packages they believed would be important in the future. Table 3 presents the software application packages that were mentioned.[6] You will notice that SAS was the most popular. The FY82 Project Leader survey indicated

---

[6] Donald P. Trees, *Results from the FY82 Project Leaders' Computing Profile Survey:* Service Satisfaction, Future Resource Requirements and Policy Issues, (Santa Monica, California: The Rand Corporation, June 1983).

Table 3

SOFTWARE APPLICATION PACKAGES OF FUTURE IMPORTANCE

Software Application Package	Number of Respondents Mentioning*	Percent of Survey Respondents	Percent of Eligible Respondents
SAS	47	41%	86%
STATLIB	34	30%	62%
SPSS	17	15%	31%
DYL 260/280	4	4%	7%
GDDM/PGF	4	4%	7%
SAS/GRAPH	3	3%	6%
BMDP	3	3%	6%
LISREL	3	3%	6%
IMSL	2	2%	4%
INFO/SYS	2	2%	4%
PROSE	1	1%	2%
MAXLIK	1	1%	2%
SIR (DBMS)	1	1%	2%
DISSPLA	1	1%	2%

* Number of survey respondents was 115. Only those respondents who indicated they could run at least one application package were considered eligible and asked to indicate applications packages of future importance to them for the open-ended question. Fifty-five (55%) of the project leaders indicated they could or did run an application package.

that the programming languages, FORTRAN, PL/I, BASIC, and C were expected to be the most important.

Data Management and Statistical Analysis

The data management of numeric data files was done by using data management packages, user written programs in programming languages such as PL/I and FORTRAN, and a combination of the two. Data management packages consist mainly of software application packages with some data management facilities, such as SAS. The statistical analysis was primarily done with statistical analysis packages. Table 4 presents the actual percentages for the methods used in data management and statistical analysis.

DATA MANAGEMENT METHODS USED AT RAND

This section will discuss the data management methods currently being used at Rand, with emphasis on large social science data bases. With the exception of one project, current projects are using data management methods other than DBMS. The reader should consider these methods, to determine if a DBMS has a place in the research environment.

As noted in section III, the social science data bases are processed in two stages: the master file stage and the analysis file stage. There are several methods being used for the data management of master files. The analysis is usually done with Statistical Software Packages (SSP); therefore the analysis files are in SSP format and use data management facilities provided by the SSP. Most of the effort is put into the data management of master files because they are more general and tend to be larger. Usually analysis files are specialized and used only for a short period of time.

There are many data management methods currently being used at Rand. These include:

Raw data files with user programs

Statistical Software Packages (SSP)

Custom software

A commercial DBMS, called Scientific Information Retrieval (SIR)

Combinations of the above methods

Table 4

DATA MANAGEMENT AND STATISTICAL ANALYSIS OF NUMERIC DATA

Data Management	
The project primarily used:	
User written programs in a high level language (PL/I, FORTRAN)	26 %
Existing data management packages (consisting mainly of application software with data management facilities)	39 %
Both	26 %
No data management tasks involved	10 %
Statistical Analysis	
The project primarily used:	
User written programs in a high level language	15 %
Existing statistical analysis packages	56 %
Both	7 %
No computerized data analysis tools	21 %

From interviewed respondents only (N=41)

Raw Data Files with User Programs

Master files usually begin in a raw data file form. A raw data file is sometimes described as being in card-image form, where the data elements are assigned a position (columns) on a card. Card (decks) files are rarely used now, but it is a good way of visualizing raw files. The user must be concerned with the physical storage of the data elements, and access the data by specifying the card or record type and the column position. The physical locations of data elements are documented in a codebook.

The actual physical organization of a raw file varies. A file can be a simple rectangular form, where there is one basic record type; or a file with several record types, with a record type variable; or a file with variable length records, with summary cards indicating the length of the record. Programming languages such as PL/I and FORTRAN are used to navigate through raw data files. Therefore, the use of raw data files requires careful attention to the physical location of the data and the knowledge of a programming language. Appendix C is an example of a typical PL/I program used on a raw file with several record types and summary cards.

Statistical Software Packages (SSP)

Many projects have their master files converted to a statistical software package (SSP) format. This is a nice arrangement because the master file and analysis file are in the same form. Other advantages are that the SSP format enables the user to access information by data element name (without having to worry about the physical location), and the SSP programming language is designed to be easier to learn and use than a standard programming language (PL/I, FORTRAN). The major drawback of a master file in a SSP format is that the file structure is restricted to flat files, which do not always accommodate the master file structure. An example is a survey which has many kinds of card types, such as marriage, births, and employment. Each card type is dependent on an event; therefore it is difficult to know how many events there are per respondent and how to combine the data. Another consideration is that the master file depends on a particular SSP. When another SSP procedure needs to be used, the file must be converted to the other SSP format.

Custom Software

Custom software is written because either a SSP or a raw data file format does not provide acceptable access or management of the data base. What is acceptable is dependent on the amount of money available for the project and the longevity of the data base. In other words, those projects which can afford software development, and believe the

advantages outweigh the development costs of the software, are likely to produce custom software.

An example of custom software is a set of utility programs written for the Housing Assistance Supply Experiment (HASE) project. Rand was responsible for collecting and preparing the data on landlords, tenants, and homeowners for the Housing and Urban Development Department. The creation of the data base involved sequential processing using PL/I with the aid of HASE utilities. The HASE utility programs included: dictionary creation, print, update, merge, creation of a PL/I structure, and creation of a SPSS (Statistical Package for the Social Sciences) structure. The major advantage of the HASE software was the dictionary, which allowed data access by name. Even though these utilities were helpful on the HASE project, they are rarely used on other project data bases.

For another example, a program called RETRO was written to process life history data of the Malaysia and Guatemala Retrospective surveys. These questionnaires had data on life histories covering marital status changes, contraceptives, pregnancies, schooling, migration, employment, and income. The records on these files are indexed by the survey respondent and an event (e.g., marriage event). The RETRO program enables the user to retrieve a data element at the time of the occurrence of an event (e.g., marital status at age 20). Without RETRO, this type of retrieval would require a fairly complicated computer program. Actually RETRO is very powerful but has a poor user interface. The retrieval language is coded, column oriented, and awkward. Debugging RETRO programs requires one to use the computer column ruler, to check command column locations. This can be tedious and puts one back in the "stone age" of computing. The documentation is difficult to interpret, adding to the difficulty of learning the system. Furthermore, the RETRO system only works on the Malaysia and Guatemala Retrospective files. Appendix D has an example of a RETRO program.

Dan Relles, a researcher at Rand, designed the PRTEXT (Print/Extract Computer System for Large Data Sets) data management system. This system reads, transforms, reorganizes, and writes out data records. The system interfaces with user-written FORTRAN or PL/I subroutines which perform variable recoding, record selection, and other

complex steps. Dan Relles described the PRTEXT data mangement system as:

> a compromise between two existing alternatives: processing data under the control of a package program (e.g. SAS, SPSS, SIR), and processing data in a programming language such as FORTRAN or PL/I. Package programs may not be totally acceptable, for reasons of flexibility and efficiency. The package designers give you a subset of some programming language, but do not include all of its features (e.g. subroutines, functions, multiply-subscripted arrays); and execution time efficiency may suffer from the code that is not hard-wired to the problem at hand. On the other extreme, performing data management totally within FORTRAN or PL/I programs is very undesirable: it is quite easy to make format statement mistakes; and formatted input/output operations are very inefficient. PRTEXT provides a way of obtaining programming language flexibility and efficiency while avoiding these input/output difficulties.[7]

A file can be produced from PRTEXT that can be directly input to STATLIB, a computing library. The designer of PRTEXT also co-wrote STATLIB. Currently, only the designer and his group are using PRTEXT. PRTEXT is still in the development stage, though there is no real incentive to continue the work. Like the RETRO system, the effort was put into getting a workable data management system, and not into the user interface and documentation. Another alternative to developing custom software is the use of a DBMS.

Commercial DBMS

There is also a commercial DBMS called SIR (Scientific Information Retrieval) available at Rand. SIR is a unique commercial DBMS because it was designed for the research market. The SIR programming language is modeled after the language used in SPSS, a popular statistical software package. Another feature is that several types of statistical software package files can be directly built from SIR. Rand was chosen as the test facility when SIR was being developed for IBM systems. Therefore, Rand did not have to pay for the software. During an

[7] Daniel A. Relles, *PRTEXT: A Print/Extract Computer System for Large Data Sets (Preliminary Documentation)*, (Santa Monica: The Rand Corporation, 1981).

experimental stage, two projects were selected to test the system. Only the programmer time was charged to the project, while the computer time was charged to the computer center.

The Malaysia project was one of the test projects. The project programmer liked aspects of SIR, in particular the hierarchical and network structures available and the fact that the unit of observation could be changed without much effort.[8] However, he felt SIR was expensive and tended to be inefficient in computer time and storage. His general feelings were good, and noted that the project would have used the product more if RETRO was not already available. RETRO is able to do most of the difficult retrievals, it is less expensive to run, and the programmer is already experienced in the use of RETRO. Therefore, even though some of the files were already in SIR form, the programmer prefers using a combination of RETRO, user written PL/I programs, and SAS.

The other project had data on criminal justice. This project wanted to use the SPSS statistical software package for analysis, but SPSS can only handle 500 variables. Researchers decided not to use SAS because they wanted a direct link to SPSS, which SIR provides. Listed below are the programmer's impressions of SIR.[9]

The direct interface to SAS, SPSS and BMDP; never having to use a "raw" file.

Good documentation for the record type, variable name and variable labels.

The variable label information was used to edit incoming data and this information is also passed to statistical package files.

The 31 levels of security are allowed and can be assigned on a variable level.

The master file was always the current file (did not have to worry about versions).

---

[8] Personal interview with Terry Fain, Senior Programmer Analyst, The Rand Corporation, 3 April 1984.

[9] Personal interview with Suzanne Polich, Manager of Application Programmer Staff, The Rand Corporation, 11 April 1984.

Good backup facilities.

Most retrievals were straight-forward.

This was a case where SIR was a good choice. It is interesting to note, that this project was done a few years ago, and only used the batch mode. SIR has since been enhanced and will be discussed further in section VII.

Combination of Data Management Methods

It is common for data management of a data base to utilize a combination of the methods described above. Since analysis is usually done through a statistical software package (SSP), each data management method must transfer data to the SSP. In the case of custom software, many are used in combination with programming languages (e.g., HASE utilities, PRTEXT). These systems are mainly used to facilitate programming language tasks. The RETRO system is an independent custom program, but is only available for two files. Other files in the Malaysia and Guatemala data bases are in raw form and accessed through programming languages.

V. BUSINESS VS RESEARCH

Data base management systems have become widely used in the business environment. DBMS have been considered the most important software in the business world. The total market for data base products exceeded $440 million in 1982. The market is expected to reach $2 billion by 1988.[1] In contrast, DBMS has not been very successful in the research world.[2] [3]

This section will compare the research environment and the business environment. This comparison will attempt to explain why commercial DBMS are successful in the business world and not in the research world. The following topics will be considered:

Type of Data/Applications

Ownership

Computer Resources

Application Programming Support

Data Base Retrieval

Data Base Accessibility

Data Base Update

Interface with Other Software

Portability

---

[1] Robert Kerin,"The Most Important Software In the Business World," *Software News*, December 1983, p. 25.

[2] E. Cohen and R. A. Hay Jr, "Why are Commercial DataBase Management Systems Rarely Used for Research Data?" in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 132.

[3] Arie Shoshani, "Statistical Databases: Characteristics, Problems and Some Solutions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, ed. James Gentle (New York: North Holland Publishing Company, 1983) p. 10.

Conclusions on DBMS Configurations

Given the three possible DBMS configurations for a research environment, the DBMS interface with SSP seems to be the most reasonable solution. There are varying degrees of interface between the DBMS and the SSP, from some programs to help the transfer of data, to a system allowing complete access to SSP procedures from the DBMS. In fact, a complete interface system could appear to the user as the optimal configuration of the DBMS containing all the data management and statistical analysis features. There are currently no complete DBMS and SSP interface systems.

IMPORTANT DBMS FEATURES SURVEY DESCRIPTION

To determine the importance of various DBMS features, the Rand computer users were asked to rate the features in two ways. The first section had the respondents rate the DBMS features independently using a graduated scale (rating method 1). Because there was a possibility that all the features might be considered desirable, an additional question was asked. The respondent had to rank the features with respect to each other, choosing the five most desirable features (rating method 2). The following fifteen features were rated (given in the same order used in the questionnaire):

(Q1) Data base flexibility (restructuring)
(the ability to change the data base structure, to add or delete types of data)

(Q2) Data entry editing
(input or update data is checked for valid values, e.g., within ranges)

(Q3) Nonredundancy of data
(controlled integration of data to avoid the inefficiency and inconsistency of duplicated data)

(Q4) Integrity of data from system failure
(reconstruction and recovery facilities)

(Q5) Security/Privacy
(prevent access to the data set, specific units of data, types of data or combinations of data)

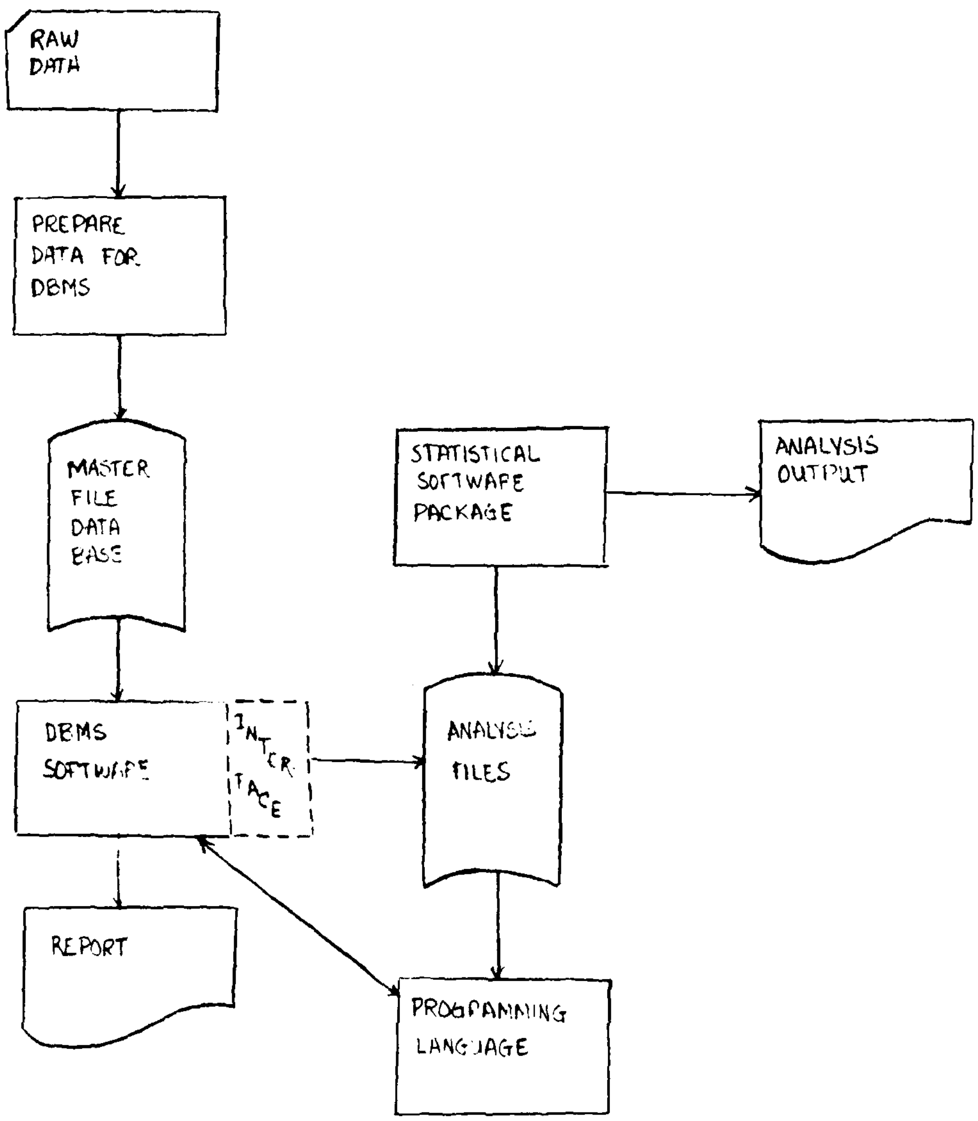


Fig. 3 -- DBMS Interface with Statistical Software Package

been patterned after the syntax of SPSS (Statistical Package for the Social Sciences), a widely used statistical system. SIR is a fairly complete DBMS, allowing hierarchical and network data structures. The U.S. Census was put in the SIR DBMS successfully.[6]

Other DBMS may have indirect interfaces with SSP, written as extensions to the DBMS. Most of these interfaces have been written by users and not the software vendors. There has been no incentive for either the DBMS vendors or the SSP vendors to provide interface software, probably because each vendor does not realize the importance of the other.[7] If there is no interface available, then the DBMS should at least provide output that is easy to change into a SSP compatible file and the SSP should accept a wide variety of input. Figure 3 shows what the DBMS and SSP interface configuration might look like.

DBMS with Statistical Procedures

Most DBMS have little or no support for statistical analysis. A DBMS which includes statistical analysis procedures would be the optimal configuration, but an unrealistic expectation. The statistical analysis requirement puts a large burden on the DBMS, though there are those taking on the challenge of developing DBMS which provide statistical facilities. These existing systems have a limited range of statistical functions, and cannot support all the necessary analysis tools.[8] Even a SSP which specializes in statistical procedures may not include all the necessary methods; sometimes more than one SSP must be used.

---

[6] Richard Hay Jr., "Data Definition in SIR/DBMS as Applied to the 1980 U.S. Census," in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), pp. 134-140.

[7] Ivor Frances and J. Sedransk, "Software Requirements for the Analysis of Surveys," *Proceedings of the 9th International Biometric Conference*, 1976 p. 235.

[8] FOCUS is a system with a fair amount of statistical procedures. *FOCUS User's Manual*, (New York: Information Builders Inc., [n.d.]).

DBMS Interface with SSP

A DBMS that interfaces with statistical software packages could be the best of both worlds, containing all the benefits of a DBMS and all the statistical analysis methods available in a SSP. We will look at two DBMS (RAPID and SIR) with a built-in interface to SSP.

Statistics Canada is a research institution which conducts and processes the major socio-economic surveys and censuses for the Canadian Federal Government. They were using the commercial DBMS TOTAL, but felt it was not efficient enough to handle the 1976 Census with 8 million households and 23 million persons. This led to the development of RAPID (Relational Access Processor for Integrated Databases).[4] RAPID was designed to manage large and complex statistical data bases on an IBM 370 compatible system. RAPID is based on the relation model, using a transposed file system. The RAPID relations are accessed through host languages (PL/I, COBOL, FORTRAN) and assembly language, passing the name of a command and other parameters. Future plans for RAPID are to develop a data query language for non-procedural access to data. Most of the usual DBMS utilities are available. There are also interfaces to statistical packages (SAS, SPSS, and P-STAT) and tabulation packages (TPL, EXTRACTO/OPTIDATA, EASYTRIEVE). The RAPID system is not available as a commercial product, but can be made available to governments, institutions such as universities and other non-profit organizations. RAPID has been implemented in statistical agencies in at least eight countries.

The other statistical DBMS is SIR (Scientific Information Retrieval).[5] SIR is unusual because it is a commercial DBMS that was specifically developed for the research market needs. SIR interfaces with several statistical software packages (SPSS, BMDP, SAS, P-STAT), producing output files in the statistical package format. Another nice feature of SIR is that the data description and retrieval language has

---

[4] M. J. Turner, R. Hammond, and P. Cotton, "A DBMS for Large Statistical Databases," *Fifth International Conference on Very Large Data Bases* (IEEE, Oct. 1979) pp. 319-327.

[5] Barry N. Robinson et al., *SIR, Scientific Information Retrieval: Version 2 User Manual Version 2*, (Evanston, IL: SIR Inc., 1980).

Given the available software there are three possible DBMS configurations [3]:

1. The DBMS and the statistical software packages maintained separately.
2. The DBMS with an interface to statistical software packages.
3. The DBMS with data management and the statistical analysis capabilities.

DBMS and SSP Maintained Separately

In this configuration, the DBMS is used to maintain and retrieve data, and the Statistical Software Package (SSP) is used for statistical analysis. When the DBMS and the statistical software package are maintained separately, the data transfer from the DBMS to the SSP must be done manually. The basic steps involved in the data transfer are: extract data from the DBMS, put the data into a SSP compatible form (in a flat file), and read the data into the SSP. The order in which these steps are performed varies depending on the DBMS's and SSP's ability to transform data. The above steps are used to build an analysis file from the master file. The master file resides in the DBMS, while the analysis file resides in the SSP. After the data transfer has been done manually several times, the natural progression is to develop utility programs to simplify the task of building analysis files. An example is a program to automatically set up the input format used by the SSP (e.g., a custom program written at Rand had automatic input formats for SSP). These utility programs are the start of an interface system between the DBMS and the statistical software package.

[3] Hideto Ikedo, Yasuyuki Kobayashi, "Additional Facilities of a Conventional DBMS to Support Interactive Statistical Analysis," in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 25.

VI. RESEARCH ENVIRONMENT DBMS REQUIREMENTS

The previous sections have described the "research environment" and some of the data management needs. Now that we have an understanding of what a research environment is, this section will discuss the interrelationship between a DBMS and the research environment. The first portion of this section presents possible DBMS configurations. The second portion is an attempt to determine the important DBMS features, via a survey of computer users in a research environment (The Rand Corporation).

DATA BASE MANAGEMENT SYSTEM CONFIGURATIONS

This section will propose how a data base management system (DBMS) can be incorporated into the research environment. Large social science data sets require software that provides data management and statistical analysis. There are two types of software products which specialize in each of these functions. A data base management system provides powerful capabilities of data base organization, retrieval and update, but rarely supports statistical analysis procedures. DBMS capabilities were discussed in detail in Section II. Statistical software packages (SSP) are capable of supplying statistical analysis procedures, but typically their data management facilities are restricted to flat files.[1] Large social science data sets do not always fit these constraints.[2] To get an understanding of the SSP capabilities, appendix E lists the characteristics of one of the popular SSP, the Statistical Analysis System (SAS).

---

[1] J. D. McKeen, C. W. Slinkman, "Statistical Analysis Support for Database Management Systems," Jane F. Gentleman ed., *Proceedings of the Computer Science and Statistics: 12th Annual Symposium on the Interface*, (Ontario, Canada: University of Waterloo, 1979), pp. 295-296.

[2] Example of where other software was needed is covered in the data management methods portion of Section IV.

3. Is there any promise for DBMS in the Research Environment?

Yes, but only for "large" and/or complex social science data bases, with reasonable funds. It is difficult to determine exactly when it would be advantageous to implement a DBMS. Research projects strive to minimize the use of computer resources, sometimes at the sacrifice of flexibilities or features. But, it could be a false economy if it takes an experienced programmer a lot of work and time to access the data. The programmers' time becomes more important as software costs become a greater and greater part of total computer costs.[16] A DBMS query language could also allow nonprogrammer staff to access the data directly.

Many of the research data management needs would also benefit the business environment. Examples are better query systems, ability to do ad hoc retrievals, and interface with other software. Some of these features are already coming along, in an effort to become more "user friendly."[17]

---

[16] Ian Palmer, *Data Base Systems: A Practical Reference* (Wellesley, Massachusetts: Q.E.D. Information Sciences Inc., 1975) pp. 5-19, 5-20.

[17] Some DBMS with some of these features are evaluated in Section VII.

several computer systems and provide facilities for efficient transfer of data bases between systems. There are DBMS which are implemented on various computer systems; the major reason was to have a larger market rather than for portability reasons. Therefore facilities for transferring data between computer systems has not been a major emphasis.

Conclusions from Business and Research Comparison

To summarize the comparison of the business and the research environments, three questions will be considered:

1. Are DBMS designed for the Business Environment?

From the success rate of DBMS in the business world, one could deduce that DBMS were designed for the corporate/business application. Whether this design practice was intentional or not, some reasons might be: there is more familiarity with business data and its applications; and on the economic side, the business market is very large.

2. Are there significant differences between Business and Research data management needs?

The comparison on the preceding pages indicated that the research environment has very different data management requirements, such as different retrieval and update specifications and the need for interface with other software. Another major difference is the amount of financial support for processing of data, in terms of computer time, storage, and programming staff. The business environment provides a vast amount of this support, while the research project has limited resources and must be concerned about computer costs. DBMS have been optimized for business requirements, which may not necessarily benefit research needs (e.g., informational retrievals, simultaneous access). The business environment can afford DBMS even if they are not totally satisfied with it, while the research project must be more selective.

Management Systems Rarely Used for Research Data?" in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 133.

Interface with Other Software

Data in a business data base is either queried for immediate use or used in the preparation of a report. Most DBMS provide a nice report writing feature, so it is not necessary to leave the DBMS system, though there are some instances where it would be nice to transfer data to other software systems (e.g., spreadsheets, graphic systems).

The purpose of the research data retrieval is to build an analysis file from the master file controlled by the DBMS. The analysis file must be statistically analyzed. In general DBMS do not provide statistical procedures. If the user is involved in statistical analysis, the user must either write his/her own procedures in a host language or use a SSP. Using SSP is the easier option even though one must be concerned with data transfer.

DBMS rarely have a direct transfer capability to other software. The DBMS data transfer capabilities, usually consist of dumping the data in a raw (card image) form. In this case all definitions and labels are lost. Data management of the analysis file is usually done within the SSP. It would be nice if the labels and definitions from the DBMS could be transferred to the SSP. The interface between DBMS and SSP is discussed in more detail in section VI.

Portability

Business/corporate data bases contain "enterprise" data which is owned by the corporation and will stay there. The business data base is rarely used by other companies. For this reason, portability to other computer systems is not a prime concern in the business environment.

Often social science data bases or other research data bases will contain information of interest to many researchers and it is desirable to install the data base at several locations on a variety of computer systems. Conducting a survey or experiment is very expensive, so it seems reasonable that more than one group should have access to the data base. Another reason for portability is that researchers frequently migrate to other academic institutions and they need to take their data bases.[15] The research environment needs a DBMS which is available on

[15] E. Cohen and R. A. Hay Jr, "Why are Commercial DataBase

few researchers feel the need to use the interactive TSO system. Interactive systems are more expensive to run, and unless there are more advantages than speed, it is not worth it. Research data base retrievals do not require instantaneous actions. The typical retrieval will extract portions of the master file to create an analysis file. As for the concurrent capability, the research application usually will not require this, because of the limited staff.

The requirements of speed and concurrency put a large burden on the DBMS. To accomplish the goal of speed, there must be interaction with, and usually modification of, the operating system.[13] Then concurrency involves an elaborate set of algorithms to assure the integrity of the data during multiple access (e.g., only one user updating the same portion of data).

Data Base Update

Many business data bases are dynamic, meaning that they are constantly being updated. For example, an inventory data base must be updated when shipments arrive and when items are sold. In contrast, the research data base is relatively static, in that once the data has been entered it will rarely be updated.[14] Updates occur only a few times, usually during initial editing. Most social science data comes from surveys and once the data is collected, there should be no major additions. Some surveys are given repetitively (longitudinal surveys), but the data is usually analyzed after all the data has been collected.

DBMS deal well with dynamic data bases providing good updating features. Since most business applications require concurrent access, the DBMS must also be concerned with the concurrent accessing of a dynamic data base. Since the research data base is basically static, it is not necessary to support the costly concurrent data update features.

---

[13] E. Cohen and R. A. Hay Jr, "Why are Commercial DataBase Management Systems Rarely Used for Research Data?" in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 133.

[14] Arie Shoshani, "Statistical Databases: Characteristics, Problems and Some Solutions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, ed. James Gentle (New York: North Holland Publishing Company, 1983) p. 15.

The "statistical research" query, on the other hand, involves few columns and many or all rows, and is a large amount of data. The researcher uses data base retrievals to build the analysis file.[9] The analysis file is then used for subsequent statistical analysis. The analysis file must be in a form that is suitable for statistical analysis. This usually involves creating a flat file.[10]

The business "informational" query works well on DBMS, by using restrictive data linkage structures. Business retrievals are usually based on particular items (keys) and the data base structures can be designed to take advantage of this. For this reason DBMS do not perform well under ad hoc queries. Research queries are typically ad hoc in nature,[11] depending on the hypothesis being tested and the data needed to support it. One can not predict what sample and variables will be selected. Even business retrievals are not always predictable, so there is a movement toward less restrictive data base arrangements in DBMS technology (e.g., the relational DBMS).

Data Base Accessibility

The business application typically requires that many users need simultaneous access to the data base. The simultaneous requirement implies both speed and concurrency. A typical example might be an airline reservation data base, where a specific set of data is needed quickly.

Unlike the business application, the research application does not usually need such rapid access to the data base.[12] At The Rand Corporation, the interactive batch system (WYLBUR) is very popular and

---

[9] An explanation of the analysis file is given in Section II.

[10] The analysis file is built into a flat file by merging, aggregating and disaggregating data from the master file.

[11] M. J. Turner, R. Hammond, and P. Cotton, "A DBMS for Large Statistical Databases," *Fifth International Conference on Very Large Data Bases* (IEEE, Oct. 1970) p. 320.

[12] Arie Shoshani, "Statistical Databases: Characteristics, Problems and Some Solutions," in *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, ed. James Gentle (New York: North Holland Publishing Company, 1983) p. 10.

Who does the programming in the research environment is determined by the amount of funds available to the project. There are times when the researcher must assume the role of both the DBA and applications programmer.[7] In cases when the researcher has some help, this individual is often a research assistant with limited programming experience and no data base design training. Experienced programmers tend to be expensive to hire; therefore they are used only when absolutely necessary.

DBMS should accommodate the non-programmer user, by providing interfaces which are easy to use in the construction and retrieval of the data base. DBMS typically have complicated data definition languages (DDL) and retrievals are done via host languages, such as FORTRAN, COBOL, and PL/I. These languages can be difficult for the research staff to learn. Even the business personnel are not entirely content with these interface languages. Hence, there is an effort to construct more "user friendly" DBMSs with better query systems. A good compromise might be to have an experienced programmer set up the data base, while the research staff access the data through a query language.

Data Base Retrieval

Many of the data base retrievals done in business applications are "informational" queries, where there are few rows and many columns.[8] In this type of retrieval, the DBMS must search for a particular item in the data base which satisfies a given condition and display all (or most) of the attribute values of the one item. The retrievals involve a small amount of data.

---

[7] E. Cohen and R. A. Hay Jr, "Why are Commercial DataBase Management Systems Rarely Used for Research Data?" in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 132.

[8] Roy Hammond, "Metadata in the RAPID DBMS," in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 129.

Computer Resources

The business environment provides a dedicated or semi-dedicated computer with unlimited access to peripheral storage. In this case, the more the computer is used, the more it justifies the cost of the computer system. In the research environment at Rand, a project must compete with other users (projects) for computer memory, disk storage and CPU time. All computer costs are charged to the research project. Therefore, the project must strive to minimize the use of computer resources.

For a DBMS to be successful in a research environment, it must have features which conserve computer resources. For example, storage space and I/O costs could be reduced if data compression techniques were provided to take advantage of social science data attributes. Some of these attributes are sparse and qualitative data. Sparse means that most of the values are blank or null, and only a few entries have discrete values; and qualitative means there are only a few discrete values (e.g., sex).[5]

Application Programming Support

In the business environment there is a data base administrator (DBA) who is responsible for the management of corporate data bases. The DBA is usually a skilled programmer who has experience in data base design and applications. In addition to the DBA, there is usually a staff of applications programmers who interact with the data base and the DBMS through a variety of host languages. Then there is the set of data base users, who do no programming and interact with the data through application programs, or a query system.[6]

---

[5] Roy Hammond, "Metadata in the RAPID DBMS," in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 129.

[6] C.J. Date, *An Introduction to Database Systems* 3rd rev. ed. (Menlo Park, California: Addison-Wesley Publishing Company, 1981), p. 6.

Type of Data/Applications

Business/corporate data bases are comprised of enterprise data. Some examples are employee records, payroll, banking information, inventory, and airline reservations. Business applications are fairly familiar to most individuals; a typical example is the generation of monthly bank account statements.

Social science data bases are not as commonly known. One that most of us are familiar with is the national census. Information is collected on population and income. This data can be studied to determine trends in family composition, housing, and employment. Social science data bases are described in more detail in Section III and Appendix A.

Ownership

Who owns the data base influences the amount of support the data base will receive. Support can be defined in terms of money, manpower, and time. The business/corporate data base belongs to the company as a whole, and is considered an enterprise resource. The social science data base on the other hand is independent and used on projects where the data is pertinent.[4] Business data has the backing of the company or corporation, while social science does not. Research projects are usually funded by research grants from the government and other sources. Therefore, the support given to a research data base is usually less than that given to a business data base. This is illustrated in the "computer resources" and "application programming support" sections below.

---

[4] However, there is usually a data facility in charge of building, maintaining,and distributing the social science data base. Rand has a data archival center which keeps account on the various data bases available at Rand.

(Q6) Self documentation of data base
(data directory/dictionary)

(Q7) Data independence
(changes to the data base structure does not require modification of programs which use the data)

(Q8) Multiple views of data
(provision for complex file structures and access paths)

(Q9) Interactive processing

(Q10) Interface with host language
(PL/I, FORTRAN)

(Q11) Simple query/report language
(instead of, or in addition to interface with host language)

(Q12) Interface with statistical software package
(the ability to produce statistical software package files directly)

(Q13) Concurrent access to data
(more than one user can access the data base at the same time)

(Q14) Standardization in facilities for storage, modification, retrieval, and reorganization (so there is a uniform method)

(Q15) Portability
(across a wide range of computers and operating systems)

The features were rated individually using this scale:

1 Very Important

2 Important

3 Neutral

4 Little Importance

5 Not Important

A copy of the questionnaire is given in Appendix F.

Samples and Response Rates

The Rand computer users were divided into three groups (samples): the members of the computer services department, members of the research staff, and members of the management information system group. Each group was mailed the DBMS questionnaire, with a different cover letter. The cover letters used are in Appendix G.

The computer services department (CSD) consists of application programmers, computer consultants, the computer systems group, and managers. The application programmers are assigned to various research projects; therefore they have a good understanding of research needs. Sixty-one questionnaires were sent within CSD, and there was a 79 percent response rate. One person did not feel qualified to respond, so the N (total number of respondents) for CSD is 47.

The research staff (RS) are either researchers or research assistants from various departments. Some of the departments are behavioral sciences, engineering and applied sciences, economics, information sciences, political science, and systems sciences. Members of this group were difficult to select, because many researchers do not do their own computer programming and may not have data management experience. Project leaders from the FY82 Project Leader Survey,[9] and researchers and research assistants with known programming experience were contacted. There was a 51 percent response rate. Five responded that they did feel qualified to respond, so the N for the research staff is 24.

The management information system (MIS) group is made up of computer programmers in the financial department, who are responsible for the corporate business programming. There are six members of the MIS programming staff. Four of these responded to the DBMS survey. Even though this is such a small sample, this group was included to get an indication of business application needs.

The overall response rate for all groups was 66 percent, which is fairly good considering the method of collection was mailed surveys and that the members of the research staff were difficult to select. The

[9] Refer to Appendix B for information on FY82 Project Leader Survey.

results of the survey will be discussed with respect to the different sample groups: the Management Information System Group (MIS), and the Research Environment (CSD and RS).

The Management Information System (MIS) Group

The average rating given to all DBMS features was 1.7, where 1 is very important and 2 is important. The MIS respondents considered (Q5) Security/Privacy, (Q7) Data Independence, and (Q11) Simple Query/Report Language as being very important. It is interesting to note that individually (Q11) Simple Query/Report Language was rated very highly, but was rarely considered as one of the top five features. Also, even though the MIS group are experienced programmers, they still want a Simple Query/Report Language. (Q1) Data Base Flexibility and (Q13) Concurrent Access to Data, were other features that were considered fairly important. The second rating method indicated that the (Q7) Data Independence and (Q1) Flexibility features were very desirable. Other important features were (Q13) Concurrent Access, (Q4) Integrity, and (Q5) Security/Privacy.

When the results from both rating systems were summarized, the features that were considered important were: (Q7) Data Independence, (Q1) Flexibility, and (Q5) Security/Privacy. The data base flexibility and data independence features are related; the MIS programmers want the ability to restructure the data base and have the current computer programs be independent of these changes. The Security/Privacy issue implies that there is sensitive data that must be protected. An example of sensitive data might be employee salary information.

The least important feature was (Q15) Portability. This is because corporate/business data rarely needs to be moved to other computer systems. The (Q12) Interface with Statistical Software Package (SSP) feature was rated neutral. There was only a limited amount of statistical analysis done, therefore it was not considered critical to interface with SSP. Both these results verify statements made in Section V, "Business Environment vs Research Environment."

The Research Environment (CSD & RS)

The Computer Service Department (CSD) and the Research Staff (RS) responses were combined to form the research environment. The CSD and RS groups had similar results. In fact the top six features for both groups are the same, with only a slight difference of order. The only feature that was rated differently was (Q14) "Standardization in Facilities" for storage, modification, retrieval and reorganization. This is probably because the CSD programmers are assigned to various research projects which usually deal with different data management methods. The reader should refer to Tables 5 through 8 for detailed feature ratings.

The top five DBMS features for the Rand research environment are:

- (Q1) flexibility of the data base
- (Q4) integrity of data from system failure
- (Q12) interface with SSP
- (Q7) data independence
- (Q6) self documentation of data base

Most of these top features are fairly familiar, except for the "interface with statistical software packages" feature. This feature is significant because a business environment might agree with the other features, but not with "interface with SSP."

The least important features were (Q15) Portability, (Q5) Security/Privacy, (Q11) Query/Report Language and (Q9) Interactive Processing. Even though the portability issue is more likely to affect researchers, they still did not consider it important.[10] One reason for the low rating of the security/privacy feature might be that Rand has a dedicated computer and set of procedures for dealing with sensitive data. Some readers might be surprised that the interactive processing

[10] Refer to Section II, for description of conceptual models.

and query/language features were not considered important. This is because the Rand researchers are not accustomed to real interactive processing; they use the WYLBUR interactive batch system.

Table 5

RATING THE IMPORTANCE OF DBMS FEATURES

Question	Feature Division	very	2	3	4	not
		Degrees of Importance				
Q1: Data Base Flexibility						
	CSD	68%	23%	9%	0%	0%
	RS	79%	17%	4%	0%	0%
	CSD & RS	72%	21%	7%	0%	0%
Q2: Data Entry Editing						
	CSD	32%	51%	11%	4%	2%
	RS	38%	33%	25%	4%	0%
	CSD & RS	34%	45%	16%	4%	1%
Q3: Nonredundancy						
	CSD	23%	43%	26%	9%	0%
	RS	17%	33%	29%	13%	8%
	CSD & RS	21%	39%	27%	10%	3%
Q4: Integrity						
	CSD	64%	15%	21%	0%	0%
	RS	58%	33%	0%	8%	0%
	CSD & RS	62%	21%	14%	3%	0%
Q5: Security/Privacy						
	CSD	17%	36%	32%	9%	6%
	RS	13%	8%	46%	17%	17%
	CSD & RS	16%	27%	37%	11%	10%
Q6: Self Documentation						
	CSD	43%	45%	11%	2%	0%
	RS	42%	29%	21%	4%	4%
	CSD & RS	42%	39%	14%	3%	1%
Q7: Data Independence						
	CSD	48%	41%	9%	2%	0%
	RS	39%	39%	22%	0%	0%
	CSD & RS	45%	41%	13%	1%	0%

Table 5 (continued)

Q8: Multiple Views						
	CSD	15%	52%	30%	2%	0%
	RS	30%	39%	17%	13%	0%
	CSD & RS	20%	48%	26%	6%	0%
Q9: Interactive Processing						
	CSD	17%	43%	28%	11%	2%
	RS	21%	21%	25%	17%	17%
	CSD & RS	18%	35%	27%	13%	7%
Q10: Interface with Host Language						
	CSD	34%	38%	23%	2%	2%
	RS	41%	32%	18%	9%	0%
	CSD & RS	36%	36%	22%	4%	1%
Q11: Query/Report Language						
	CSD	19%	40%	36%	4%	0%
	RS	14%	18%	50%	14%	5%
	CSD & RS	17%	33%	41%	7%	1%
Q12: Interface with Statistical Software Package						
	CSD	40%	43%	15%	2%	0%
	RS	50%	33%	13%	0%	4%
	CSD & RS	44%	39%	14%	1%	1%
Q13: Concurrent Access						
	CSD	45%	30%	13%	6%	6%
	RS	8%	46%	29%	13%	4%
	CSD & RS	32%	35%	18%	8%	6%
Q14: Standardization in Facilities						
	CSD	36%	43%	11%	8%	2%
	RS	13%	48%	26%	4%	7%
	CSD & RS	29%	44%	16%	7%	4%
Q15: Portability						
	CSD	15%	32%	34%	17%	2%
	RS	0%	25%	41%	29%	4%
	CSD & RS	10%	30%	37%	21%	3%

Table 6

TOP RANKED DBMS FEATURES

THE COMPUTER SERVICES DEPARTMENT ONLY

question / ranked number:	1	2	3	4	5	total
Q1: Data Base Flexibility	16	5	6	4	3	34
Q4: Integrity	10	9	6	3	3	31
Q7: Data Independence	4	5	5	5	6	25
Q12: Interface with SSP	2	3	4	5	8	22
Q6: Self Documentation	3	4	5	5	4	21
Q2: Data Entry Editing	1	6	2	4	5	18
Q14: Standardization	1	3	2	4	6	16
Q13: Concurrent Access	0	2	3	6	3	14
Q3: Nonredundancy	1	6	3	2	0	12
Q10: Interface with Host Lang.	3	2	1	4	1	11
Q9: Interactive Processing	3	1	0	3	1	8
Q8: Multiple Views	1	0	4	1	2	8
Q5: Security/Privacy	1	1	2	0	2	6
Q11: Query/Report Language	0	0	2	1	2	5
Q15: Portability	1	0	2	0	1	4

Table 7

TOP RANKED DBMS FEATURES

THE RESEARCH STAFF ONLY

question / ranked number:	1	2	3	4	5	total
Q1: Data Base Flexibility	12	3	1	0	3	19
Q4: Integrity	2	3	4	4	2	15
Q6: Self Documentation	1	2	3	3	4	13
Q12: Interface with SSP	4	3	2	1	2	12
Q2: Data Entry Editing	4	1	2	2	1	10
Q7: Data Independence	0	2	3	3	1	9
Q8: Multiple Views	0	2	1	0	5	8
Q10: Interface with Host Lang.	0	2	3	1	1	7
Q9: Interactive Processing	0	1	0	4	0	5
Q11: Query/Report Language	0	1	2	1	1	5
Q14: Standardization	0	0	0	2	1	3
Q3: Nonredundancy	0	2	1	0	0	3
Q5: Security/Privacy	0	1	1	1	0	3
Q13: Concurrent Access	0	0	0	1	1	2
Q15: Portability	0	0	0	0	0	0

Table 8

TOP RANKED DBMS FEATURES

THE COMPUTER SERVICES AND RESEARCH STAFF

question / ranked number:	1	2	3	4	5	total
Q1: Data Base Flexibility	28	8	7	4	6	53
Q4: Integrity	12	12	10	7	5	46
Q6: Self Documentation	4	6	8	8	8	34
Q7: Data Independence	4	7	8	8	7	34
Q12: Interface with SSP	6	6	6	6	10	34
Q2: Data Entry Editing	5	7	4	6	6	28
Q14: Standardization	1	3	2	6	7	19
Q10: Interface with Host Lang.	3	4	4	5	2	18
Q8: Multiple Views	1	2	5	1	7	16
Q13: Concurrent Access	0	2	3	7	4	16
Q3: Nonredundancy	1	8	4	2	0	15
Q9: Interactive Processing	3	2	0	7	1	13
Q11: Query/Report Language	0	1	4	2	3	10
Q5: Security/Privacy	1	2	3	1	2	9
Q15: Portability	1	0	2	0	1	4

VII. EVALUATION OF COMMERCIAL DATA BASE MANAGEMENT SYSTEMS

Section V examined the differences between business and research data management needs. In Section VI, the possible DBMS configurations and the important DBMS requirements of a research environment were discussed. Information from both these sections, was used to evaluate what commercial DBMS have to offer the research environment.

A set of seven Data Base Management Systems was selected for evaluation based on the following criteria: the DBMS or vendor is established and well known, the set of DBMS must include a variety of conceptual models (hierarchical, network, and relational,[1] and the set must include DBMS which are designed for research applications (e.g., the SIR DBMS). Other DBMS could have been included, but the DBMS included in Table 9 are fairly representative.

Table 9

COMMERCIAL DBMS EVALUATED FOR USE IN A RESEARCH ENVIRONMENT

DBMS	*Vendor*
IMS/VS (Information Management System/VS)	IBM
IDMS & IDMS/R	Cullinet Software
TOTAL	Cincom Systems
ADABAS (Adaptable Data Base System)	Software AG of North America, Inc.
SIR (Scientific Information Retrieval)	Scientific Information Retrieval, Inc.
ORACLE	Oracle Corporation
SQL/DS (Structured Query Language Data System)	IBM

[1] Refer to Section II, for description of conceptual models.

Table 10 is a chart which compares the features of these seven DBMS. The major references were the recent Data Decisions DBMS survey [2] and interviews with DBMS vendors. A vendor warned me that any DBMS literature and reviews might be out of date because the industry is constantly upgrading.[3] The evaluation compares the DBMS listed above, with emphasis on the DBMS features considered important by Rand computer users [4] and some of the features discussed in the comparison between research and business requirements.[5] These issues were evaluated: data base flexibility, data base retrieval, data base recovery, data documentation, data base security, portability, interface to statistical software packages, purchase cost and computer resources.

Data Base Flexibility

The data base flexibility feature is the ability to easily change/restructure the data base. The ADABAS system has a high level of flexibility because the data is organized by inverted lists and does not have a schema. The ORACLE and SQL/DS are relational systems, which are organized by tables and therefore have a flexible structure. These DBMS are flexible because of their structural design, while the hierarchical or network design is more rigid. SIR is a hierarchical/network system which has an automatic data base restructuring facility. The restructure command mechanism is easy, but the data base must be reloaded, which could be a costly procedure.

---

[2] *Data Decisions Software Product Evaluation: DBMS*, (Cherry Hill, New Jersey: Data Decisions, 1983).

[3] Telephone interview with Kurt Fainman, Branch Manager, Oracle Corporation, 7 May 1984.

[4] The most important features were: flexibility of the data base, integrity of data from system failure, interface with SSP, data independence, and self documentation of data base.

[5] Refer to Section V.

System	IMS/VS (Information Management System/VS)	IDMS (Integrated Database Management System), IDMS/R	Total
Vendor	IBM	Cullinet Software	Cincom Systems Inc.
Number of Installations	over 3,000	over 1,400	5,500
Computers/Operating Systems	IBM 370, 3000, 4300, and compatible	IBM 370, 3000, 4300, and compatible; all mainframe operating systems	All of the mainframe-vendor systems and most of the mini-vendor systems
Minimum Memory Requirements	192k bytes & up	200K - 259K bytes	14K to 60K
Data Base Type (Organization)	hierarchical (sequential and direct)	hierarchical/network with inverted-list capabilities IDMS/R CODASYL & relational	hierarchical/network with inverted list capabilities
Query Facilities	SQL/Data System	On-Line Query (OLQ) On-Line English (OLE)	TASK simple query language (optional)
Host Language Interface	COBOL, PL/1, assembler, RPG II for DOS/VS	COBOL, FORTRAN, PL/1, RPG-II, assembler, language with CALL	CMS, COBOL, FORTRAN, Host assembler, RPGII, IBM cards
Statistical Software Package Interface	No	No	No
Data Base Security	terminal ID and password verification, keyword access	password protection, database procedures and subschema	password protection to field level
Report Generator	SQL/Data System	CULPRIT	SOCRATES (batch)
Data Entry/Editing	no	?	interactive ENV-DATA
Easy Data Base Restructure	no	seems relatively good	no
Data Dictionary	DB/DC	Integrated Data Dictionary (IDD)	Data Control System (DCS) provided separately

System	ADABAS (Adaptable Data Base System)	SIR (Scientific Information Retrieval)	ORACLE
Vendor	Software AG of North America	Scientific Information Retrieval, Inc.	Oracle Corporation
Number of Installations	1,000	?	over 350 worldwide
Computers/Operating Systems	IBM 360/40, 370/125, 4300, 30, SSX/VSE, DOS, DOS/VS, DOS/VSE, OS/MVT, OS/MFT, OS/VS1, OS/VS2,(SVS), OS/VS2(MVS), and CMS component of VM/370; VAX 11/780, and VAX 11/750 Siemens 4004/45; PBS, BS100, BS2000	CDC CYBER series, NOS, NOS/BE, IBM 360/370, OS/VS, DEC VAX ; VMS. Honeywell CP-6. PERKIN-ELMER, PRIME, SIEMENS, UNIVAC	DEC PDP-11/23 through 11/70, RSX-11M, IAS, RSTS/E, and UNIX. DEC VAX-11; VAX/VMS, VAX/UNIX. Harris Mini-computers, VULCAN. Data General, AOS/VS. IBM 370 computers; VM/CMS. Motorola M68000 microcomputer; Xenix
Minimum Memory Requirements	220K bytes - 750K bytes	?	100K bytes
Data Base Type (organization)	inverted list with quasi-relational features	hierarchical and network	relational
Query Facilities	ADASCRIPT integrated, NATURAL (optional)	SPSS like language, SIR/SQL+ (optional)	SQL
Host Language Interface	any high level language with CALL facility	FORTRAN and other host languages thru SIR/HOST	scheduled precompiler for COBOL and FORTRAN
Statistical Software Package Interface	SAS/Graph, therefore probably SAS	Descriptive Statistical procedures. Automatic creation of SPSS, SAS, P-STAT and BMDP	available overseas
Data Base Security	password for update, at file, field, and record level; encryption	at field and record level	password to system, owner-controlled access to files on tables, views, and fields.
Report Generator	ADACOM (batch)	integrated report generator	ORATOR
Data Entry/Editing	No	SIR/FORMS (optional for data entry)	integrated full screen
Easy Data Base Restructure	*yes, because of inverted list structure	yes (modify schema command)	flexible
Data Dictionary	PREDICT (integrated)	yes (IDD)	integrated provided separately

The ORACLE and SQL/DS are both relational DBMS. The relational structure is represented in the form of tables. Tables are the natural way of viewing data for researchers, because the data is represented as tables in SSP. The relational structure also lends itself to research (ad hoc) queries, because this structure does not rely on predefined paths. The Structured Query Language (SQL) has a very good reputation for being easy to use and well documented. SQL is available in SQL/DS, ORACLE and now even in SIR. A good query language could increase programmers' productivity and give the research staff the ability to do their own programming. The DBMS relational architecture is relatively new. These commercial DBMS have not been proven yet, but have great promise. One advantage of a new DBMS is that the vendors are more willing to try or add other features. For example, the ORACLE vendor seemed interested in developing interfaces to SSP.

Can DBMS succeed in a research environment?

The major factors in the success of DBMS in a research environment are: selection of the appropriate DBMS, full support from management, and selection of the appropriate data bases to be implemented on the DBMS. Selecting the right DBMS is not an easy task and involves extensive evaluation with consideration of the features available, cost, and ease of use. SIR, ORACLE and SQL/DS are good DBMS candidates.[12] Management plays a key role in the success of the DBMS in a corporation. The use of the DBMS must be promoted; consulting facilities and classes must be provided. The DBMS should not be used for all data bases, because some data bases are not appropriate and would not benefit.[13] For example, SAS should still be used for some data bases.

What is the future of DBMS for use in the research environment?

Cullinet Software, 8 May 1984. Telephone interview with Kurt Fainman, Branch Manager, Oracle Corporation, 7 May 1984.

[12] Refer to Section VII for more information on these DBMS.

[13] Criteria for data bases that would benefit from DBMS was given in Section VIII.

To get the opinions of a research community, a survey was designed and taken at The Rand Corporation. Computer users were asked to rate the importance of DBMS features.[9] The five most important features are listed here.

data base flexibility--the ability to change the data base structure.

integrity--reconstruction and recovery facilities.

interface to statistical software packages (SSP)--the ability to produce statistical software package files directly.

data independence--the ability to make changes to the data base structures without requiring modification of programs which use the data.

self documentation of data--data directory/dictionary facilities.

Most of these features are fairly standard, with the exception of the "interface to SSP" feature which is unique to the research environment. The business environment would rarely use a SSP.

What are the results of an evaluation of commercial DBMS for use in a research environment?

The commercial DBMS that were evaluated included: IMS/VS, IDMS, TOTAL, ADABAS, SIR, ORACLE, and SQL/DS.[10] The DBMS which are good candidates for use in a research environment are SIR, ORACLE and SQL/DS. They are also the least costly in terms of license fees.

Of the DBMS evaluated, SIR (Scientific Information Retrieval) was the only one which was especially marketed for the research or scientific applications. SIR provides an interface to several statistical software packages (SSP). When most DBMS vendors were asked about interface to SSP, they knew of some users who built their own interface.[11]

---

[9] The survey and results were discussed in Section IV. A copy of the survey questionnaire is in Appendix F.

[10] Refer to Section VII and Tables 9, 10, and 11 for details on these DBMS.

[11] Telephone interview with Cherie Guenfeld, IDMS representative,

are supported within projects from research grants. Research projects must be more conservative and selective when choosing software products, because they can not afford to pay for a system that does not fulfill their needs. Furthermore, many DBMS are optimized for business applications, which are not always useful to research applications. The main reason DBMS have not been successful in the research environment is because they only provide a limited number of advantages, and therefore are not worth the implementation cost.

How would a DBMS fit into the research environment?

As stated in Section VI the research data base is processed in two stages: the master file stage and the analysis file stage. The "master file" is the original data put into a computer readable form; the "analysis file" is built from master files and is the actual file used in analysis.[8] The DBMS would be used with master files. These files tend to be large, complex, have a generalized use, and are stable. The analysis files are usually small, in flat file form, specialized and used for a short period. Statistical analysis is performed on analysis files. Statistical Software Packages (SSP) provide statistical procedures and data management facilities adequate for most analysis files. Given the DBMS and SSP software, the possible configurations for a DBMS in a research environment were discussed in Section VI. At present most research environments which implement DBMS maintain the DBMS and SSP separately. Because the analysis file is built from the master file, it would be useful if there were an interface between the DBMS and SSP. In that way, information such as labels and valid value ranges could be transferred from the DBMS to the SSP directly. The DBMS and SSP complement each other, and the interface configuration provides the advantages of both software products.

What DBMS features are considered important to the research community?

[8] Refer to Section III for a detailed description on the master file and the analysis file stages.

The research/statistics community has some justification for claiming that current DBMS are not geared toward their needs. When examining the DBMS features, it seems as though most DBMS are designed for business applications. One reason for this design direction is that the business market is much larger than the research market. One DBMS vendor did not believe it was cost effective to go after the research market.[3] Another reason might be that the business environment is much more widely known than the research environment.[4] DBMS were designed with no real conception of the research application.

What are the differences between the business environment and the research environment?

Given that DBMS are geared toward business applications, are research demands significantly different? Most DBMS vendors believe their product is general enough to be used by other applications.[5] Researchers, on the other hand, believe they have unique needs. In Section V of this report, the research environment and business environment were compared. It appears that the business and research applications do have different demands on a DBMS. Some of the differences are that the research applications need to use statistical analysis procedures, that data retrievals tend to be ad hoc in nature, and that data is fairly static, requiring only periodic updates.[6]

Actually, the most significant difference is the amount of support provided for the data base, where data base support is defined in terms of computer time, storage cost, and programming staff.[7] Business data bases have the backing of the entire company, while research data bases

---

[3] Telephone interview with Monte B. Hipple II, ADABAS representative, AG Software, 7 May 1984.

[4] This is why two sections (III, IV) of this report were dedicated to describing the research environment.

[5] Telephone interview with Monte B. Hipple II, ADABAS representative, AG Software, 7 May 1984.

[6] Refer to Section V for more details.

[7] Section V compares the ownership, computer resources, and application programming support for business and research.

IX. SUMMARY AND CONCLUSIONS

Most researchers feel that current commercial data base management systems (DBMS) are not suited to their needs,[1] even though DBMS have been successful and accepted by the corporate/business community.[2] Most computer scientists do not believe there are significant differences between the data base management needs of research applications and business applications. I include myself in this group. As a programmer analyst at the Rand Corporation, I became curious about why a data base management system (DBMS) was not being used. This was the motivation for the investigation of DBMS software for use in a research environment. The results are presented in this section. The following questions will be answered.

Are DBMS primarily designed for business applications?

What are the differences between the business environment and the research environment?

How would a DBMS fit into the research environment?

What DBMS features are considered important to the research community?

What are the results of an evaluation of commercial DBMS for use in a research environment?

Can DBMS succeed in a research environment?

What is the future of DBMS for use in the research environment?

Are DBMS primarily designed for business applications?

---

[1] E. Cohen and R. A. Hay Jr, "Why are Commercial DataBase Management Systems Rarely Used for Research Data?" in *Proceedings of the First LBL Workshop on Statistical Database Management*, ed. Harry K. T. Wong (Berkeley: Lawrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 132.

[2] Robert Kerin, "The Most Important Software In the Business World," *Software News*, December 1983, p. 25.

Research projects which currently employ other data management methods are less likely to convert.[10] Most research projects last a few years; therefore as projects are initiated, they might be more open to using a DBMS. This provides a good schedule for converting to DBMS in stages.

---

[10] An example was given in Section IV, when the Malaysia project programmer chose to remain with their current software, rather than convert to the SIR DBMS.

1. The programmer no longer maintains data files. The file is initially organized by the data base administrator and maintained by the DBMS.
2. The data base is internally documented by a data dictionary, which maintains descriptions of data and relationships among the data.
3. The programmer can access the data base through logical views created in subschemas. The programmer is no longer restricted to the physical location formats.
4. There is flexibility in altering the initial data base structures to fulfill new needs. Modifications can be made with little effort and cost.
5. Simple query systems increase programmers' productivity. In addition, research assistants, who currently avoid programming because of its complexity, may feel encouraged to do more programming. The research assistant's time is usually less expensive than a programmer's time.

There are other considerations that may encourage the use of a DBMS, such as centralized control, uniformity in data access, good data documentation facilities, and flexibility.[8] A careful analysis of DBMS must be made in order to choose the appropriate one.[9]

Transition to DBMS

The computer management group plays a key role in converting over to a DBMS. Once they have accepted the DBMS as a viable solution, they must provide support for the product and encourage its use. Support would include: announcing the product availability in a computing bulletin (these are regularly published for Rand computer users), give courses onsite (Rand computer services staff could teach the courses), and provide consulting (also available from the computer services staff).

---

[8] DBMS features were discussed in more depth in Section II.
[9] The selection of DBMS was considered in Section VII.

- large and complicated data bases (with many types of records)
- single data base used to address many diverse hypotheses (e.g., psychologist, economist, statistician all using the data base to answer different questions).
- data bases in use for a long period (years)
- data bases which will be added to over time
- data bases where costs are shared with other research projects or other sites (a distributed data base)

Arguments for the use of a DBMS at Rand

If Rand is to consider a DBMS, the DBMS must be easier to use and more cost efficient than SAS and the user written data management software.[4] These two requirements can be satisfied by an appropriate DBMS. This is not easily substantiated, because one must consider the several factors. Some of these factors are computer resources cost verses programmer cost, programmer productivity, and the other DBMS features (advantages).

As noted before, the research environment is more concerned with minimizing costs than the business environment.[5] On the surface, the added cost of product license, computer machine time and storage would inhibit the acceptance of a DBMS. Yet, if a more in depth study is done, a DBMS would appear more favorable. Although a generalized DBMS is less computer resource efficient than a custom high level language program, the trade off in machine efficiency is compensated by the programmers' reduced time in program development and maintenance.

When DBMS are used with appropriate data bases, the application program development is flexible, faster, and less expensive than current methods.[6] DBMS increase software productivity in the following ways [7]:

---

[4] The DBMS would be used with data bases which fit the criteria described above.

[5] Refer to Section V for differences between the research and business environments.

[6] Andrew B. Whinston and C. W. Holsapple, "DBMS for Micros," *Datamation*, April 1981, p. 165.

[7] For numbers 1-4 the reference was Micheal Gagle and Gary J. Koehler, "Data Base Management Systems: Powerful Newcomers to Microcomputers," *Byte*, November 1981, p. 122.

VIII. INTRODUCTION OF DBMS INTO A RESEARCH ENVIRONMENT

This section briefly discusses how a DBMS might be introduced into a research institution. The social science data bases at the Rand Corporation are used as a case study. Guidelines are recommended as to the type of data bases that should be incorporated in the proposed DBMS. Arguments are given for the use of a DBMS at Rand and finally, some suggestions are made for the transition to a DBMS.

Data Bases for use in a DBMS

Currently at Rand, the SAS statistical software package is the major application software product being used. In addition to the statistical analysis features, SAS has a substantial set of data management facilities.[1] These data management facilities are sufficient for data bases which have simple structures and/or are relatively small in size. When more complicated data bases are used, other methods must supplement SAS data management facilities. Typically PL/I and FORTRAN programs are written to supplement SAS's deficiencies. The programs are either direct file handling programs or custom utility programs.[2] Data Base Management Systems (DBMS) could be used for these complex data bases.

For the DBMS to be successful, the data bases implemented must be prescreened. Suzanne Polich, manager of the applications programmer staff at Rand, suggested some criteria for selecting those social science data bases which would benefit most from DBMS [3]:

[1] Refer to Appendix E, which describes SAS features.

[2] For more details, refer to "The Data Management Methods Used at Rand" portion of Section IV.

[3] Note from Suzanne M. Polich, Manager of Application Programmer Staff, The Rand Corporation, 7 May 1984.

Table 11

COMMERCIAL DBMS: (Dollars) FIVE YEAR LICENSE FEE COMPARISON

DBMS	(Dollars) *Five Year License Purchase*
IMS/VS	92.1K - 361.3K
IDMS & IDMS/R	117K - 279K
TOTAL	61.9K - 149.6K
ADABAS	106K - 230K
SIR	20K - 60K
ORACLE	12K - 96K
SQL/DS	22.6K - 39.4K

As for the computer resource usage, this is even more difficult to judge without benchmarking and running performance analysis. One indicator of computer resource usage is the Minimum Memory Requirements entry in Table 10. TOTAL has a modest requirement of 60K bytes, compared to the other systems where 250K bytes is about average. The DBMS requiring the most memory was IBM SQL/DS with 2M bytes. A large memory requirement is not necessarily a characteristic of relational DBMS because ORACLE uses 100K bytes.

As expected, the SIR DBMS is best suited to the research environment. Its one major advantage over other DBMS is the interface to several of the major SSP. Next in order are the relational DBMS, in particular SQL/DS and ORACLE. The clear advantage to the relational design is the table orientation, which researchers are accustomed to, because SSP require tabular data. Also the Structured Query Language seems to be easy to learn and use, so the researcher can do most of his/her own programming. The relational design is the direction DBMS are going, and they seem suited to the research environment.

Interface to Statistical Software Packages

Because of the big demand for statistical analysis in the research environment, an interface from DBMS to statistical software packages (SSP) would be useful. The SIR system has interfaces to the SAS, SPSS, P-STAT, and BMDP statistical software packages. As mentioned throughout this paper, SIR is a unique commercial DBMS, in that it is pursuing the research application market. This is why SIR has the best SSP interface. Interface to SSP is usually not included in DBMS literature, so vendors were contacted for more information. ADABAS has an interface to SAS/Graph, a SAS Institute graphics package, which is compatible with the statistical analysis software.[14] The ORACLE vendor knew of an SSP interface system being used overseas, but did not know what particular SSP.[15] The Cullinet vendor knows of many IDMS users that interfaced to SAS, but was not aware of a direct interface.[16] The IDMS report writer gives the user the ability to easily extract data in any form. The IBM and Cincom Systems vendors did not know of any SSP interfaces available.[17] [18]

Cost Comparison

It is difficult to make comparisons about DBMS cost, because initial pricing is dependent on the machine, operating system, and the optional software included. To get a general idea of the DBMS costs, a price list is given in Table 11.[19] [20]

---

[14] Telephone interview with Monte B. Hipple II, ADABAS representative, AG Software, 7 May 1984.

[15] Telephone interview with Kurt Fainman, Branch Manager, Oracle Corporation, 7 May 1984.

[16] Telephone interview with Cherie Guenfeld, IDMS representative, Cullinet Software, 8 May 1984.

[17] Telephone interview with Marti Operman, DBMS representative, IBM, 7 May 1984. Telephone interview with Marie Myer, TOTAL representative, Cincom Systems, 7 May 1984.

[18] Telephone interview with Marie Myer, TOTAL representative, Cincom Systems, 7 May 1984.

[19] *Data Decisions Software Product Evaluation: DBMS*, (Cherry Hill, New Jersey: Data Decisions, 1983).

[20] *Computerworld Buyers Guide*, (Framingham, Mass.: CW Communications Inc., December 1983)

making the query languages simple and user friendly (e.g., the ADABAS NATURAL system). The IBM Structured Query Language (SQL) [9] has a good reputation in the industry, good enough that other vendors have borrowed the language for their systems (e.g., ORACLE and SIR). It is believed that the SQL/DS system is easy enough and documented well enough for a user to learn in one day.[10]

Portability

To get an indication of the DBMS portability across computers, refer to the Computer/Operating Systems entry in Table 9. IDMS, IMS/VS, and SQL/DS seem to be restricted to IBM machines, though Data Decisions mentioned that SQL/DS documentation implied that it could run on "alien machines."[11] The other DBMS run on a much larger variety of computers. Another interesting development is that the Cullinet Software and Oracle Corporation have built DBMS software for the IBM PC. Cullinet provides communication between the IBM PC and mainframe IDMS/R.[12]

There is also the issue of portability across software. Cullinet IDMS provided migration aids which transfer data from other DBMS (e.g., from IBM DL/1 to IDMS). In this way, the vendor makes it easier to transfer to his system. Portions of the data may also need to be transferred to other software for further processing. The ADABAS system interfaces to at least 30 software products, including SAS/Graph and IBM GDDM.[13]

---

[9] C.J. Date, *An Introduction to Database Systems* 3rd rev. ed. (Menlo Park, California: Addison-Wesley Publishing Company, 1981), pp. 97-105.

[10] Gabrielle Wiorkowski, "Relational DBMS Meets the Real World," *Data Management*, Sept. 1983, p. 36.

[11] *Data Decisions Software Product Evaluation: DBMS*, (Cherry Hill, New Jersey: Data Decisions, 1983).

[12] *Data Decisions Software Product Evaluation: DBMS*, (Cherry Hill, New Jersey: Data Decisions, 1983).

[13] Telephone interview with Monte B. Hipple II, ADABAS representative, AG Software, 7 May 1984.

Data Base Retrievals

The data base structure or type also has an effect on the data base retrievals. As mentioned in Section V, research queries/retrievals are typically ad hoc in nature, making it difficult to determine predefined paths or structures. Inverted list DBMS and relational DBMS handle ad hoc retrievals quite nicely. This is not true for most hierarchical/network systems, such as IMS/VS and TOTAL. Hierarchical/network systems work well with production programs, which use the same path repetitively. Conversely, relational DBMS perform poorly on production retrievals. To speed production retrieval, the SQL/DS system allows the user to build predefined access paths into indexes that surround the tables.[6]

Data Base Recovery, Data Documentation and Data Base Security

All the DBMS evaluated had facilities for data base recovery, data documentation, and data base security. The recovery systems were usually very good with check point restart and transaction logging. Data documentation was provided in terms of a data dictionary. The major difference among systems was whether the dictionary was integrated or optional. One example of a fully integrated data dictionary is ADABAS which uses the data dictionary to bypass the use of a data definition language. The data base security features consisted of protecting the data from access and update. The protection could be put on a range of levels (e.g., file, record, and field). Some systems had encryption available.

Query Languages

There are various query languages available; some DBMS use a procedural type language, while others can be quite graphic, like IBM "Query By Example."[7], [8] Many of the DBMS vendors have put an effort into

---

[6] *Data Decisions Software Product Evaluation: DBMS*, (Cherry Hill, New Jersey: Data Decisions, 1983).

[7] C.J. Date, *An Introduction to Database Systems* 3rd rev. ed. (Menlo Park, California: Addison-Wesley Publishing Company, 1981), pp. 181-202.

[8] James A. Larson, "DBMS Develops a Human Personality," *Data Management*, Sept. 1983, p. 38.

Table 10 (continued)

System	SQL/DS (Structured Query Language Data System
Vendor	IBM
Number of Installations	unknown
Computers/Operating Systems	IBM 370 Models 138, 145, 148, 158; 4331, 4341; 3031, 3033s; IBM DOS/VSE AF, SSX, VM/SP
Minimum Memory Requirements	2 mega-bytes
Data Base Type (organization)	relational
Query Facilities	SQL, Query by Example Interactive SQL (ISQL)
Host Language Interface	COBOL, PL/I, FORTRAN, assembler
Statistical Software Package Interface	No
Data Base Security	password; system authority level to view and field, inter-user privilege grants/revokes
Report Generator	Interactive SQL (ISQL)
Data Entry/Editing	ISQL
Easy Data Base Restructure	very flexible
Data Dictionary	integrated

The future is very promising. Many of the features needed by the research environment are being developed, such as the ability to do ad hoc queries, and "easy to use" query languages. Relational DBMS are proliferating; they seem well suited to research needs. It is also important to note the availability of DBMS on microcomputers; now the price is more affordable. Initially these DBMS had a limited number of features, but now full DBMS such as ORACLE are being implemented.[14] Data analysis can be done on a microcomputer because statistical software packages, such as ABSTAT, Microstat, Statpac, and SYSTAT are also available.[15] Microcomputer software should be sufficient for small social science data bases and some analysis files. The master file can be on a mainframe DBMS (such as ORACLE and IDMS/R) and data can be down-loaded directly to the microcomputer DBMS.

There is still the issue of interface between DBMS and SSP. Software designers for each of these products feel their product is self sufficient and independent. If an interface is needed to the other software product, neither vendor wants to provide it. The DBMS vendors believe the SSP should provide the interface and the SSP vendors believe the DBMS should provide the interface. Both these groups need to work together. There may be legal difficulties in interfacing to other products.

The DBMS and SSP interface problem is an example of the need for more communication between the computer science community and the research/statistics community.[16] There are some efforts being made toward this goal, such as the Computer Science and Statistics Symposium on the Interface[17] and the Lawrence Berkeley Laboratory Workshop on

---

[14] Telephone interview with Kurt Fainman, Branch Manager, Oracle Corporation, 7 May 1984.

[15] James Carpenter, "Statistical Software for Microcomputers," *Byte, April 1984, p. 235.*

[16] Where the computer science community is defined as the software designers and developers, and the statistics community is made up of statisticians and other researchers who develop and use statistical methods.

[17] James E. Gentle, ed., *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface,* (New York: North Holland Publishing Company, 1983).

Statistical Database Management.[18] In fact, work being done at Lawrence Berkeley Laboratory goes beyond the DBMS configurations proposed in this report. They are working on a DBMS that will manage data in the master file stage and also the analysis file stage. The analysis file is usually made in several steps, such as extracting data from several files, aggregating data, and building derived variables using statistical procedures. The LBL DBMS designers are tackling the difficult problem of keeping track of the analysis process.

With the prospect of microcomputers, relational DBMS and specialized statistical DBMS, there are many possibilities for the future of DBMS in the research environment.

---

[18] Harry K. T. Wong, ed., *Proceedings of the First LBL Workshop on Statistical Database Management*, (Berkeley: Lawrence Berkeley Laboratory, University of California LBL-13851, 1982).

X. GLOSSARY

AGGREGATION - The unit is lower and must be summarized.

ANALYSIS FILE - The actual file used in analysis. This file is built from the master file.

CLEANING DATA - Updating dirty data with corrections to errors.

CONCEPTUAL MODEL - A representation of the information content of the data base.

DATA BASE MANAGEMENT SYSTEM (DBMS) - A set of programs that function to create and update the data base, retrieve data and generate reports from the data base.

DATA MODEL - Same as conceptual model.

DATA MODIFICATION - The process of modifying data elements (variables). The data element is recoded or derived from other variables.

DATA TRANSFORMATION - The process of transforming the record structure. Steps in transformation include: merging, aggregation and disaggregation.

DERIVED VARIABLE - To build a new variable by manipulating one or more variables.

DIRTY DATA - A collection of data that has errors.

DISAGGREGATION - The unit is higher and must be apportioned down.

FLAT FILE - A file which is in a data matrix form (rectangular form), where every entity is crossed with every attribute.

HIERARCHICAL MODEL - The conceptual model is in the form of a tree, where an individual record (node) may have only one owner (parent).

MASTER FILE - The original data put into a computer readable form.

MISSING VALUE - When a data value was mistakenly omitted.

NETWORK MODEL - The conceptual model is represented as a plex (graph). Relationships between fields are established via links.

RECTANGULAR FILE - Same as flat file.

RELATIONAL MODEL - The conceptual model is represented as a table.

STATISTICAL SOFTWARE PACKAGE (SSP) - A set of programs that specialize in providing statistical procedures.

UNIT OF ANALYSIS - The level the researcher is concerned with. Unit of analysis is analogous to the record key of the analysis file.

WORKING FILE - Same as analysis file.

APPENDIX A

EXAMPLE OF SOCIAL SCIENCE DATA BASES

U.S. Census Public Use Sample (PUS). The census data provides an excellent source of comparative data for studying changes in the U.S. lifestyles, family composition, and other demographic characteristics. The data sets contain information at both the household level and the individual level.

The Parnes National Longitudinal Surveys of Young Women, Young Men, Mature Women and Mature Men. Each of these surveys was administered to approximately 5,000 individuals in a U.S. national probability sample. The panels, which began in the late 1960s, have been reinterviewed during most of the following years, providing a longitudinal view of around 15 years length. The surveys focus on labor market experience, and also include data on family composition and background, marriage, childbearing, marital status changes, living arrangements, attitudes toward sex roles, employment characteristics, and family experiences.

The Health Insurance Study (HIS). This data base was developed at Rand to examine how health insurance affects the utilization of health services and the health status of individuals. The sample includes data on 7706 individuals in 2754 families in six U.S. areas. Data was gathered between 1974 and 1981 at the household, family, and individual levels. The data included information on price of health care, and a rich variety of measures of individual's physical, social, and mental health.

INCAP-Rand Guatemala Survey (IRGS) and the INCAP Longitudinal Study (ILS). These two data sets represent, respectively, retrospective and panel data on Guatemala family life. The IRGS is a socioeconomic survey administered to 1097 households in six Guatemala villages, four rural and two semi-urban, between 1974 and 1976. The IRGS includes detailed female and male retrospective life histories; time budget information; a survey of income and wealth; information on children, parents, and grandchildren; and community variables. The ILS supplements the IRGS; the ILS sample covers all children born between 1969 and 1973 in the

IRGS four rural villages. These children were followed until 1977 or age seven, whichever came first. There were 1623 children in the ILS sample. The data was on children's nutrition and health, including periodic anthropometry data for both measures for mothers and children.

APPENDIX B

DESCRIPTION OF FY82 PROJECT LEADERS' COMPUTING PROFILE

The "FY82 Project Leaders' Computing Profile Survey" was conducted to study computer utilization at the Rand Corporation.[1] FY82 refers to fiscal year 1982. Donald P. Trees of the Rand Computer Services department, designed the survey for use in evaluating cost reduction strategies and in developing future computing configurations. Data was collected in the following areas: user and use characteristics, satisfaction with hardware, software and services, future resource satisfaction with requirements, and attitudes toward possible future policy areas.

There were two types of data collection, personal interview and mail survey. The personal interview had more questions on current computer usage. The project leaders were chosen for personal interview on the basis of FY82 computing expenditures and project program area (department). Personally interviewed respondents accounted for eighty-six percent of FY82 revenues. They were considered the large and medium users. The small users were sent the mail survey. Of the 183 project leaders contacted, a total of 115 responded, which was a sixty-three percent response rate. But the personal interview response rate was ninety-three percent, with 41 respondents. The survey was biased toward large and medium users.

---

[1] Donald P. Trees, *Results from the FY82 Project Leaders' Computing Profile Survey:* Service Satisfaction, Future Resource Requirements and Policy Issues, (Santa Monica: The Rand Corporation, Computer Service Department), June 1983.

APPENDIX C

EXAMPLE OF PL/I PROGRAM USED FOR DATA MANAGEMENT

The following programs were used on the Malaysia data base. The first program is a utility program that reads a tape file onto disk. The second program is a PL/I program which extracts data from a Malaysia file. Make note of the formats in the data structures; the programmer must take great care that these are accurate. The third program is a SAS program that produces the frequencies on the data in the new file.

First Program (IBM Utility Program)

```
//M6969CPY JOB (3379,50,218),'COPY TAPE FILE',CLASS=S
//COPY1    EXEC  PGM=IEBGENER,REGION=220K
//SYSIN    DD  DUMMY
//SYSPRINT DD  SYSOUT=A
//SYSUT1   DD DSN=ROUND1.MF1,
//         DISP=(OLD,PASS),LABEL=5,
//         UNIT=(HIGH9,DEFER),
//         VOL=SER=005345,
//         DCB=(BUFNO=1,BLKSIZE=3600,LRECL=80,RECFM=FB)
//SYSUT2   DD DSN=M.M6969.A3379.ROUND1.MF1,
//         DISP=(NEW,CATLG,DELETE),VOL=SER=TEMP10,UNIT=TEMP,
//         SPACE=(TRK,(50,50),RLSE),
//         DCB=(RECFM=FB,LRECL=80,BLKSIZE=3600,BUFNO=1)
```

Second Program (PL/I Program)

```
//M6969PNM JOB (3379,0100,218),'EXTRACT FROM MF1',CLASS=N
// EXEC PLIXCG
//PLI.SYSIN DD *
*    PROCESS OF,NUMBER,SEQUENCE(73,76);
 /*BUILD FILE FROM MF1.                                           */
OMF1FILE:  PROC OPTIONS (MAIN) REORDER;
0/****************************************************************/
 /*                                                               */
 /*       INPUT:                                                  */
 /*                                                               */
 /*            MF1  HOUSEHOLD ROSTER FILE                         */
 /*                                                               */
 /*                                                               */
 /*       OUTPUT:                                                 */
 /*                                                               */
 /*            ONE RECORD PER PERSON OVER THE AGE OF 5            */
 /*                                                               */
 /*                                                               */
 /*       PURPOSE:                                                */
```

```
 /*                                                                   */
 /*       THIS PROGRAM EXTRACTS INFORMATION FROM THE MF1 FILE         */
 /*                                                                   */
 /*********************************************************************/
1/*********************************************************************/
 /*                                                                   */
 /*                    FILE DECLARATIONS                              */
 /*                                                                   */
 /*********************************************************************/
0         /*        INPUT       */
          DCL       MF1IN       FILE RECORD  INPUT SEQUENTIAL;
0         /*        OUTPUT      */
0         DCL       WFILE      FILE RECORD OUTPUT SEQUENTIAL;
          DCL       SYSPRINT   FILE STREAM OUTPUT PRINT;
0/*********************************************************************/
 /*                                                                   */
 /*                   WORKING STORAGE VARIABLES                       */
 /*                                                                   */
 /*********************************************************************/
ODCL EOFMF1          BIT(1),
     MF1IN_CNTR      FLOAT DEC,
     OUT_CNTR        FLOAT DEC,
     CASEID_CNTR     FLOAT DEC,
     BLANK           CHAR(1) INIT(' '),
     (YES,NO)        BIT(1);
ODCL REC_PTR         POINTER;
ODCL (NSKIP,CNT,I) FIXED BIN(15);
ODCL ONLOC           BUILTIN;
1/*********************************************************************/
 /*                                                                   */
 /*                  INPUT OUTPUT AREAS                               */
 /*                                                                   */
 /*********************************************************************/
ODCL 1 MF1_SMRY , /*MF1 SUMMARY CARD STRUCTURE*/
       2 FILLER1     CHAR(2),
       2 SCRD_TYPE   PIC'9',
       2 SCASEID     PIC'9999',
       2 FILLER2     CHAR(4),
       2 AREA        CHAR(6),
       2 FILLER3     CHAR(30),
       2 TOTAL_CRD   PIC'99',
       2 FILLER4     CHAR(31);
0/*MF1 HOUSEHOLD ROSTER STRUCTURE*/
 DCL 1 MF1_STRUC BASED(REC_PTR),
       2 FILLER1     CHAR(2),
       2 EVENT_TYPE PIC'9',
       2 CASE        PIC'9999',
       2 FILLER2     CHAR(3),
       2 PERSID      PIC'999',
       2 FILLER3     CHAR(17),
       2 SEX         CHAR(1),
       2 RACE        CHAR(1),
```

```
      2 FILLER4     CHAR(1),
      2 AGE         PIC'99',
      2 FILLER5     CHAR(4),
      2 EDSTAT      CHAR(1),
      2 GRADE       CHAR(2),
      2 PRE54       CHAR(1),
      2 CERT        CHAR(1),
      2 MEDIUM      CHAR(1),
      2 SCHTYPE     CHAR(1),
      2 FILLER6     CHAR(34);
ODCL 1 MF1_OUT, /*CHILD INFORMATION RECORD*/
      2 CASE        PIC'9999',
      2 AREA        CHAR(6),
      2 PERSID      CHAR(3),
      2 SEX         CHAR(1),
      2 RACE        CHAR(1),
      2 AGE         CHAR(2),
      2 EDSTAT      CHAR(1),
      2 GRADE       CHAR(2),
      2 PRE54       CHAR(1),
      2 CERT        CHAR(1),
      2 MEDIUM      CHAR(1),
      2 SCHTYPE     CHAR(1);
1/*****************************************************************/
 /*                                                               */
 /*                  PROGRAM INITIALIZATION                       */
 /*                                                               */
 /*****************************************************************/
0 ON ENDFILE (MF1IN) BEGIN;
     EOFMF1 = YES;
     END;
 ON ERROR
    BEGIN;
     PUT LIST(ONLOC);
     PUT DATA;
     END;

0YES = '1'B;
 NO  = '0'B;
 EOFMF1=NO;
 MF1IN_CNTR, OUT_CNTR, CASEID_CNTR=0;

 OPEN FILE (MF1IN),
   FILE(WFILE),  FILE (SYSPRINT);

0PUT EDIT('OUTPUT FROM MF1 EXTRACT PROGRAM')
         (SKIP(3),A);
0/*****************************************************************/
 /*                                                               */
 /*               MAIN PROGRAM LOGIC                              */
 /*                                                               */
 /*****************************************************************/
```

```
0READ FILE(MF1IN) INTO(MF1_SMRY);
 IF  EOFMF1 THEN MF1IN_CNTR=MF1IN_CNTR+1;
 CASEID_CNTR=1;
0DO WHILE ( EOFMF1);
1/*********************************************************************/
 /*                                                                   */
 /*                 READ IN MF1 EVENT RECORDS.  PUT INFORMATION       */
 /*                 INTO OUTPUT RECORD.                               */
 /*                                                                   */
 /*********************************************************************/
 IF TOTAL_CRD  = 0
   THEN DO;
     MF1_OUT.AREA=MF1_SMRY.AREA;
     DO I=1 TO TOTAL_CRD;
       CALL READ_MF1;
       IF MF1_STRUC.AGE>=5
         THEN DO;
           MF1_OUT=MF1_STRUC, BY NAME;
           CALL WRITE_OUTPUT;
           END;
       END;
     END; /*TOTAL_CRD  =0*/
0/*READ NEXT CASE SUMMARY CARDS*/
 READ FILE(MF1IN) INTO(MF1_SMRY);
 IF SCRD_TYPE  =0 THEN PUT EDIT('COUNT OFF, CASEID=',SCASEID,
   '  TYPE NO=',SCRD_TYPE)(SKIP(2),A,A,A,A);
 IF  EOFMF1
   THEN DO;
     MF1IN_CNTR=MF1IN_CNTR+1;
     CASEID_CNTR=CASEID_CNTR+1;
     IF SCASEID=683 THEN TOTAL_CRD=3;
     /*ERROR ON RAW DATA(TOTAL_CRD=2)*/
     END;
0END; /*WHILE ( EOFMF1)*/
1/*********************************************************************/
 /*                                                                   */
 /*                 END OF PROGRAM'S MAIN LOGIC                       */
 /*                                                                   */
 /*********************************************************************/
0/* END OF FILE  */
 PUT EDIT
    ('MF1 IN COUNT   ',MF1IN_CNTR,
     'CASEID COUNT   ',CASEID_CNTR,
     'WFILE COUNT    ',OUT_CNTR)
    (SKIP,A,F(5));
 RETURN;
1/*********************************************************************/
 /*                                                                   */
 /*              READ MF1IN FILE                                      */
 /*                                                                   */
 /*********************************************************************/
0READ_MF1:  PROC;
```

```
    READ FILE (MF1IN) SET (REC_PTR);
    IF  EOFMF1 THEN
       MF1IN_CNTR = MF1IN_CNTR + 1;
    RETURN;
    END READ_MF1;
0/**************************************************************************/
 /*                                                                        */
 /*             OUTPUT CASEID AND AVERAGES                                 */
 /*                                                                        */
 /**************************************************************************/
 WRITE_OUTPUT:  PROC;
    WRITE FILE (WFILE) FROM (MF1_OUT);
    OUT_CNTR = OUT_CNTR + 1;
    RETURN;
 END;
0END MF1FILE;
//GO.MF1IN  DD  DSN=M.M6969.A3379.ROUND1.MF1,DISP=(OLD,KEEP),
//   VOL=SER=TEMP10,UNIT=TEMP,
//   DCB=(RECFM=FB,LRECL=80,BLKSIZE=3600,BUFNO=1)
//GO.WFILE  DD  DSN=M.M6969.A3379.ROSTER.EXTRACT,
//   DISP=(NEW,CATLG,DELETE),
//   VOL=SER=TEMP10,UNIT=TEMP,SPACE=(TRK,(15,10),RLSE),
//   DCB=(RECFM=FB,LRECL=24,BLKSIZE=3192,BUFNO=1)
//
```

Third Program (SAS Program)

```
//M6969SAS JOB (3379,49,218,20,,,2),'ROSTER VARIABLES',CLASS=N
//STEP1 EXEC SAS79,REGION=300K,OPTIONS='S=72,NONEWS,GEN=1'
//FILE1  DD  DSN=M.M6969.A3379.ROSTER.EXTRACT,DISP=OLD,
//      UNIT=TEMP,VOL=SER=TEMP10
//SAVE   DD  DSN=M.M6969.A3358.RND1.SAS.ROSTER,UNIT=USER,
//      VOL=SER=RAND32,DISP=(NEW,CATLG,DELETE),
//      SPACE=(TRK,(40,10),RLSE)
//SYSIN  DD  *
DATA SAVE.ROSTER;
INFILE FILE1;
INPUT (CASE AREA PERSID SEX RACE AGE EDSTAT GRADE PRE54
  MEDIUM SCHTYPE)(4. 6. 3. 2*1. 2. 1. 2. 4*1.);
PROC CONTENTS POSITION;
PROC FREQ; TABLES
  AREA PERSID SEX RACE AGE EDSTAT GRADE PRE54 CERT MEDIUM SCHTYPE
  ;
TITLE1 ROSTER VARIABLES;
```

APPENDIX D

EXAMPLE OF A RETRO PROGRAM

Below is an example of a RETRO program, given by the author of the RETRO software system.[1]

Multiple Records per Household from the
Malaysian Female Retrospective

```
//          JOB
//MF2  EXEC PGM=RETROALL,REGION=150K
//STEPLIB DD DSN=B.B9910.A1789.IVALIB,DISP=SHR
//GO.FT05F001 DD DDNAME=SYSIN
//GO.FT06F001 DD SYSOUT=A
//GO.FT07F001 DD SYSOUT=B
//GO.FT08F001 DD ---->   MF2 CONVERTed Input Data Location
//GO.FT09F001 DD ---->   Location for Output File
//GO.SYSIN DD *                                     Var. #
RC            1  5       T MF2T    1                run card
RD   040GET        0                                record def.
OV01004                 AREA ID     T                     (1)
OV03                    AGE         5  4097               (2)
OV04                    FIRSTPRG    T50040 1    1  OT     (3)
OV08 40   1   1         #LIVEBIR    1  10               T (4)
                                    5  4097
OV09 57991   OT         CONTRA      1  10               T (5)
                                    5  4097

/*
```

The options on the Run Control card specify that one Record Definition card and five Retrieval cards are included, that summary statistics are to be calculated, that the input data will be from MF2, and that one household is to be printed. The Record Definition card states that an output record is to be created for every nonzero occurrence of the pregnancy outcome variable. The five output variables are as follows:

(1) The area reference number (to be defined only once per household).
(2) Respondent's age in months at this pregnancy.
(3) Standard month date of first nonzero pregnancy outcome.
(4) Number of nonzero pregnancy outcomes between age 10 and this pregnancy.
(5) Most recent contraceptive type used between age 10 and this pregnancy.

---

[1] Iva Maclennan, *Retro: A Computer Program for Processing Life History Data*, (Santa Monica, California: The Rand Corporation, R-2363-AID/RF, March 1978).

APPENDIX E
THE STATISTICAL ANALYSIS SYSTEM (SAS)

The basic Statistical Analysis System (SAS) provides information orage and retrieval, data modification and programming, statistical alysis, and report writing.[1] *Information storage and retrieval atures* allow the SAS user to read data in any form from most media. e data is then organized in a SAS data set which allows data access by riable name, and is self-documenting, containing descriptions for each ıriable. *Data modification and programming* features make data ›dification easy through powerful SAS statements and functions ncluding a subset of PL/I statements). *Report writing* features allow ıe SAS user to design and produce reports in any form. The *statistical ıalysis* procedures provided by SAS range from simple descriptive .atistics to complex multivariate techniques. SAS has many features r *file handling* which include copying, editing, subsetting, ncatenating, merging, and updating data sets.

A SAS user summarized SAS characteristics as follows [2]:

Runs on IBM 360/370 and compatible machines under
OS, OS/VS, VM/CMS, DOS/VSE, and TSO.

Batch or interactive modes.

End-user friendly.

Sequential files.

Program interface with Fortran, PL/I, and Assembler.

Static or dynamic data.

Numerical or character data.

---

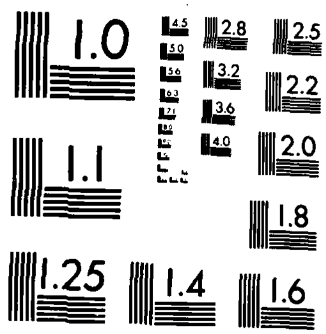
[1] *SAS User's Guide: Basic 1982 Edition*, (Cary, North Carolina: SAS nstitute Inc.), 1982.

[2] Robert H. Montgomery et al., "A Reservoir Water Quality DataBase anagement System," in *Proceedings of the First LBL Workshop on tatistical Database Management*, ed. Harry K. T. Wong (Berkeley: awrence Berkeley Laboratory, Univ. of California, LBL-13851, 1982), p. 38.

AD-A154 862 INVESTIGATION OF DBMS (DATA BASE MANAGEMENT SYSTEMS) FOR USE IN A RESEARCH ENVIRONMENT(U) RAND CORP SANTA MONICA CA P N ROSENFELD FEB 85 RAND/P-7002 2/2

UNCLASSIFIED F/G 9/2 NL

END
FILMED
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

Statistical analysis capabilities.

Complex mathematical capabilities.

Report generation.

Graphics capabilities.

Good documentation.

Technical support available.

Continuing system enhancements.

APPENDIX F
SURVEY OF THE IMPORTANCE OF DATA MANAGEMENT FEATURES

This survey is being conducted in order to identify data management features which are critical to the Rand computer user. You are asked to evaluate data base management system features that you would want if Rand were to acquire a data base management system.

Please rank the importance of each of the following data management features. Use the codes described and circle your answer.
Keep in mind that you are dealing with large data sets with many variables or many observations, or both.

When you have completed the survey, return to Pilar Rosenfeld/CSD3.

Ranking Codes

1 Very Important
2 Important
3 Neutral
4 Little Importance
5 Not Important

Features

1. Data base flexibility (restructuring) (the ability to change the data base structure, to add or delete types of data)	1	2	3	4	5
2. Data entry editing (input or update data is checked for valid values e.g. within ranges)	1	2	3	4	5
3. Nonredundancy of data (controlled integration of data to avoid the inefficiency and inconsistency of duplicated data)	1	2	3	4	5
4. Integrity of data from system failure (reconstruction and recovery facilities)	1	2	3	4	5
5. Security/Privacy (prevent access to the data set, specific units of data, types of data or combinations of data)	1	2	3	4	5
6. Self documentation of data base (data directory/dictionary)	1	2	3	4	5

7.	Data independence (changes to the data base structure does not require modification of programs which use the data)	1	2	3	4	5
8.	Multiple views of data (provision for complex file structures and access paths)	1	2	3	4	5
9.	Interactive processing	1	2	3	4	5
10.	Interface with host language (PL/I, FORTRAN)	1	2	3	4	5
11.	Simple query/report language (instead of, or in addition to interface with host language)	1	2	3	4	5
12.	Interface with statistical software package (the ability to produce statistical software package files directly)	1	2	3	4	5
13.	Concurrent access of data (more than one user can access the data base at the same time)	1	2	3	4	5
14.	Standardization in facilities for storage, modification, retrieval, and reorganization (so there is a uniform method)	1	2	3	4	5
15.	Portability (across a wide range of computers and operating systems)	1	2	3	4	5

16. Of the fifteen features just described, list the question number of the top five features, in order of importance. ____________________

17. Do you have any other comments on data management facilities for use at Rand?

____

____

____

____

____

____

____

APPENDIX G

COVER LETTERS USED WITH DBMS SURVEY

February 3, 1984

Dear MIS Programmer/Member of CSD,

I am in the process of writing a masters thesis on the feasibility of using a data base management system (DBMS) for large data sets in a research environment. As part of my report, I am using the Rand Corporation as a case study of a research environment. In order to determine the data management requirements at Rand, I am conducting a survey of computer users. The results of the survey will be made available to the Rand community. I would greatly appreciate it if you would take a few minutes to complete the attached questionnaire.

If you have any questions, please contact me at ext. 7680 (ext. 444 for messages). I would like to take this opportunity to thank you in advance for your cooperation and support.

Sincerely,

Pilar N. Rosenfeld

February 8, 1984

Dear Member of the Research Staff,

I am in the process of writing a masters thesis on the feasibility of using a data base management system (DBMS) for large data sets in a research environment. As part of my report, I am using the Rand Corporation as a case study of a research environment. In order to determine the data management requirements at Rand, I am conducting a survey of computer users.

The computer users included in the survey are the CSD and MIS programmers, and the research staff members. It is difficult to identify researchers who are concerned with data management issues. I have started by contacting project leaders involved in the CSD FY82 Project Leaders' Computing Profile survey. If you have any suggestions of who else I should contact, I would appreciate it. The results of the survey will be made available to the Rand community.

It should only take a few minutes to complete the attached questionnaire. If you have any questions, please contact me at ext. 7680 (ext. 444 for messages). I would like to take this opportunity to thank you in advance for your cooperation and support.

Sincerely,

Pilar N. Rosenfeld
Computer Services Dept.

BIBLIOGRAPHY

Anderson, Ronald E. and Francis M. Sims. "Data Management and Statistical Analysis in Social Science Computing." *American Behavioral Scientist*, Vol. 20 No. 3, January/February 1977, pp.367-409.

Appleton, Daniel S. "Implementing Data Management." *National Computer Conference Proceedings*, Vol. 49, 1980, pp. 314-316.

Carpenter, James. "Statistical Software for Microcomputers," *Byte*, April 1984, pp. 234-236, 252-264.

Computerworld Buyers Guide. Framingham, Mass.: CW Communications, Inc., December 14, 1983.

Data Decisions Software Product Evaluation: DBMS. Cherry Hill, New Jersey: Data Decisions, 1983.

Date, C. J. *An Introduction to Database Systems* 3rd rev. ed. Menlo Park, California: Addison-Wesley Publishing Company, 1981.

Domestic Research at Rand. Santa Monica, California: The Rand Corporation, January 1981.

Enger, Norman L. "Developing Data Base Structure Specifications." *Data Management*, February 1983, pp. 14-19.

Fain, Terry. Personal interview. 3 April 1984.

Fainman, Kurt. Telephone interview. 7 May 1984.

FOCUS User's Manual. New York: Information Builders Inc., [n.d.].

Frances, Ivor and J. Sedransk. "Software Requirements for the Analysis of Surveys." *Proceedings of the 9th International Biometric Conference*, 1976, pp. 228-253.

Fry, J. P. and E. H. Sibley. "Evolution of Data Base Management Systems." *Computing Surveys*, Vol. 8, March 1976, pp. 8-35.

Gagle, Michael, and Gary J. Koehler, "Data Base Management Systems: Powerful Newcomers to Microcomputers." *Byte*, November 1981, pp. 97-122.

Gentle, James E., ed. *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*, New York: North Holland Publishing Company, 1983.

Guenfeld, Cherie. Telephone interview. 8 May 1984.

Hipple II, Monte B. Telephone interview. 7 May 1984.

Kerin, Robert. "The Most Important Software In the Business World." *Software News*, December 1983, pp. 25-27.

Larson, James A. "DBMS Develops a Human Personality." *Data Management*, Sept. 1983, pp. 38-39, 68.

Maclennan, Iva. *Retro: A Computer Program for Processing Life History Data*. Santa Monica, California: The Rand Corporation, R-2363-AID/RF, March 1978.

Martin, James. *Computer Data Base Organization* 2nd rev.ed. Englewood Cliffs, New Jersey: Prentice-Hall, 1977.

McKeen, J. D. and C. W. Slinkman, "Statistical Analysis Support for Database Management Systems," Jane F. Gentleman, ed., *Proceedings of the Computer Science and Statistics: 12th Annual Symposium on the Interface*, Ontario, Canada: University of Waterloo, 1979, pp. 294-298.

Omar, Mohammed H. "DBMS Simplified." *Data Management*, Oct. 1982, pp. 23-26, 40.

Palmer, Ian. *Data Base Systems: A Practical Reference*. Wellesley, Massachusetts: Q.E.D. Information Sciences Inc., 1975.

Polich, Suzanne. Personal interview. 11 April 1984.

Rand Annual Report 1982-1983. Santa Monica, California: The Rand Corporation, 1983.

Relles, Daniel A., *PRTEXT: A Print/Extract Computer System for Large Data Sets (Preliminary Documentation)*. Santa Monica, California: The Rand Corporation, 1981.

Robinson, Barry N., et al. *SIR, Scientific Information Retrieval: Version 2 User Manual Version 2.* Evanston, IL: SIR Inc., 1980.

SAS User's Guide: Basic 1982 Edition. Cary, North Carolina: SAS Institute Inc., 1982.

Trees, Donald P. *Results from the FY82 Project Leaders' Computing Profile Survey:* Service Satisfaction, Future Resource Requirements and Policy Issues. Santa Monica, California: The Rand Corporation, June 1983.

Turner, M. J., R. Hammond, and P. Cotton. "A DBMS for Large Statistical Databases," *Fifth International Conference on Very Large Data Bases*. IEEE, Oct. 1979, pp. 319-327.

Waite, Linda J. *Changing Living Arrangements and Family Formation of Young Adults*. Proposal submitted to National Institute of Child Health and Human Development, NICHD-SBS-81-3, February 1981.

Whinston, Andrew B. and C. W. Holsapple. "DBMS for Micros." *Datamation*, April 1981, pp. 165-167.

Wiorkowski, Gabrielle. "Relational DBMS Meets the Real World." *Data Management*, Sept. 1983, pp. 36-37, 68.

Wong, Harry K. T., ed., *Proceedings of the First LBL Workshop on Statistical Database Management*. Berkeley: Lawrence Berkeley Laboratory, University of California, LBL-13851, 1982.

END

FILMED

7-85

DTIC